Chapter

# 7

# INTERVAL ESTIMATION



## 7.1 CONFIDENCE INTERVALS FOR MEANS

Given a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution $N(\mu, \sigma^2)$, we shall now consider the closeness of $\overline{X}$, the unbiased estimator of $\mu$, to the unknown mean $\mu$. To do this, we use the error structure (distribution) of $\overline{X}$, namely, that $\overline{X}$ is $N(\mu, \sigma^2/n)$ (see Corollary 5.5-1), to construct what is called a confidence interval for the unknown parameter $\mu$ when the variance $\sigma^2$ is known. For the probability $1-\alpha$, we can find a number $z_{\alpha/2}$ from Table V in Appendix B such that

$$P\left(-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1-\alpha.$$

For example, if $1-\alpha = 0.95$, then $z_{\alpha/2} = z_{0.025} = 1.96$, and if $1-\alpha = 0.90$, then $z_{\alpha/2} = z_{0.05} = 1.645$. Now, recalling that $\sigma > 0$, we see that the following inequalities are equivalent:

$$-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2},$$

$$-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le \overline{X}-\mu \le z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),$$

$$-\overline{X}-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le -\mu \le -\overline{X}+z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),$$

$$\overline{X}+z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \ge \mu \ge \overline{X}-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right).$$

Thus, since the probability of the first of these is $1-\alpha$, the probability of the last must also be $1-\alpha$, because the latter is true if and only if the former is true. That is, we have

$$P\left[\overline{X}-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\le\ \mu\ \le\overline{X}+z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]=1-\alpha.$$

So the probability that the random interval

$$\left[\overline{X}-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),\ \overline{X}+z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

includes the unknown mean $\mu$ is $1-\alpha$.

Once the sample is observed and the sample mean computed to equal $\bar{x}$, the interval $[\bar{x}-z_{\alpha/2}(\sigma/\sqrt{n}),\ \bar{x}+z_{\alpha/2}(\sigma/\sqrt{n})]$ becomes known. Since the probability that the random interval covers $\mu$ before the sample is drawn is equal to $1-\alpha$, we now call the computed interval, $\bar{x}\pm z_{\alpha/2}(\sigma/\sqrt{n})$ (for brevity), a $100(1-\alpha)\%$ **confidence interval** for the unknown mean $\mu$. For example, $\bar{x}\pm1.96(\sigma/\sqrt{n})$ is a 95% confidence interval for $\mu$. The number $100(1-\alpha)\%$, or equivalently, $1-\alpha$, is called the **confidence coefficient**.

We see that the confidence interval for $\mu$ is centered at the point estimate $\bar{x}$ and is completed by subtracting and adding the quantity $z_{\alpha/2}(\sigma/\sqrt{n})$. Note that as $n$ increases, $z_{\alpha/2}(\sigma/\sqrt{n})$ decreases, resulting in a shorter confidence interval with the same confidence coefficient $1-\alpha$. A shorter confidence interval gives a more precise estimate of $\mu$, regardless of the confidence we have in the estimate of $\mu$. Statisticians who are not restricted by time, money, effort, or the availability of observations can obviously make the confidence interval as short as they like by increasing the sample size $n$. For a fixed sample size $n$, the length of the confidence interval can also be shortened by decreasing the confidence coefficient $1-\alpha$. But if this is done, we achieve a shorter confidence interval at the expense of losing some confidence.

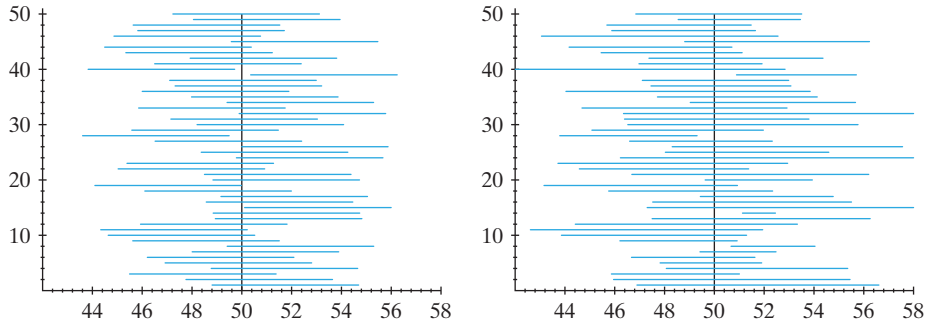
**Example 7.1-1** Let $X$ equal the length of life of a 60-watt light bulb marketed by a certain manufacturer. Assume that the distribution of $X$ is $N(\mu,1296)$. If a random sample of $n=27$ bulbs is tested until they burn out, yielding a sample mean of $\bar{x}=1478$ hours, then a 95% confidence interval for $\mu$ is

$$\begin{aligned}\left[\bar{x}-z_{0.025}\left(\frac{\sigma}{\sqrt{n}}\right),\ \bar{x}+z_{0.025}\left(\frac{\sigma}{\sqrt{n}}\right)\right]&=\left[1478-1.96\left(\frac{36}{\sqrt{27}}\right),\ 1478+1.96\left(\frac{36}{\sqrt{27}}\right)\right]\\&=[1478-13.58,\ 1478+13.58]\\&=[1464.42,\ 1491.58].\end{aligned}$$

The next example will help to give a better intuitive feeling for the interpretation of a confidence interval.

**Example 7.1-2** Let $\bar{x}$ be the observed sample mean of five observations of a random sample from the normal distribution $N(\mu,16)$. A 90% confidence interval for the unknown mean $\mu$ is

$$\left[\bar{x}-1.645\sqrt{\frac{16}{5}},\ \bar{x}+1.645\sqrt{\frac{16}{5}}\ \right].$$




(a) 90% confidence intervals, $\sigma$ known (b) 90% confidence intervals, $\sigma$ unknown

**Figure 7.1-1** Confidence intervals using $z$ and $t$

For a particular sample, this interval either does or does not contain the mean $\mu$. However, if many such intervals were calculated, about 90% of them should contain the mean $\mu$. Fifty random samples of size 5 from the normal distribution $N(50, 16)$ were simulated on a computer. A 90% confidence interval was calculated for each random sample, as if the mean were unknown. Figure 7.1-1(a) depicts each of these 50 intervals as a line segment. Note that 45 (or 90%) of them contain the mean, $\mu = 50$. In other simulations of 50 confidence intervals, the number of 90% confidence intervals containing the mean could be larger or smaller. [In fact, if $W$ is a random variable that counts the number of 90% confidence intervals containing the mean, then the distribution of $W$ is $b(50, 0.90)$.] ■

If we cannot assume that the distribution from which the sample arose is normal, we can still obtain an approximate confidence interval for $\mu$. By the central limit theorem, provided that $n$ is large enough, the ratio $(\overline{X} - \mu)/(\sigma/\sqrt{n})$ has the approximate normal distribution $N(0, 1)$ when the underlying distribution is not normal. In this case,

$$P\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha,$$

and

$$\left[\overline{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \overline{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

is an approximate $100(1 - \alpha)\%$ confidence interval for $\mu$.

The closeness of the approximate probability $1 - \alpha$ to the exact probability depends on both the underlying distribution and the sample size. When the underlying distribution is unimodal (has only one mode), symmetric, and continuous, the approximation is usually quite good even for small $n$, such as $n = 5$. As the underlying distribution becomes "less normal" (i.e., badly skewed or discrete), a larger sample size might be required to keep a reasonably accurate approximation. But, in almost all cases, an $n$ of at least 30 is usually adequate.

**Example 7.1-3** Let $X$ equal the amount of orange juice (in grams per day) consumed by an American. Suppose it is known that the standard deviation of $X$ is $\sigma = 96$. To estimate the mean $\mu$ of $X$, an orange growers' association took a random sample of

$n = 576$ Americans and found that they consumed, on the average, $\bar{x} = 133$ grams of orange juice per day. Thus, an approximate 90% confidence interval for $\mu$ is

$$133 \pm 1.645\left(\frac{96}{\sqrt{576}}\right), \qquad \text{or} \qquad [133 - 6.58,\ 133 + 6.58] = [126.42,\ 139.58]. \quad \blacksquare$$

If $\sigma^2$ is unknown and the sample size $n$ is 30 or greater, we shall use the fact that the ratio $(\overline{X} - \mu)/(S/\sqrt{n})$ has an approximate normal distribution $N(0, 1)$. This statement is true whether or not the underlying distribution is normal. However, if the underlying distribution is badly skewed or contaminated with occasional outliers, most statisticians would prefer to have a larger sample size—say, 50 or more—and even that might not produce good results. After this next example, we consider what to do when $n$ is small.

**Example 7.1-4**

Lake Macatawa, an inlet lake on the east side of Lake Michigan, is divided into an east basin and a west basin. To measure the effect on the lake of salting city streets in the winter, students took 32 samples of water from the west basin and measured the amount of sodium in parts per million in order to make a statistical inference about the unknown mean $\mu$. They obtained the following data:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 13.0 | 18.5 | 16.4 | 14.8 | 19.4 | 17.3 | 23.2 | 24.9 |
| 20.8 | 19.3 | 18.8 | 23.1 | 15.2 | 19.9 | 19.1 | 18.1 |
| 25.1 | 16.8 | 20.4 | 17.4 | 25.2 | 23.1 | 15.3 | 19.4 |
| 16.0 | 21.7 | 15.2 | 21.3 | 21.5 | 16.8 | 15.6 | 17.6 |

For these data, $\bar{x} = 19.07$ and $s^2 = 10.60$. Thus, an approximate 95% confidence interval for $\mu$ is

$$\bar{x} \pm 1.96\left(\frac{s}{\sqrt{n}}\right), \qquad \text{or} \qquad 19.07 \pm 1.96\sqrt{\frac{10.60}{32}}, \qquad \text{or} \qquad [17.94,\ 20.20]. \quad \blacksquare$$

So we have found a confidence interval for the mean $\mu$ of a normal distribution, assuming that the value of the standard deviation $\sigma$ is known or assuming that $\sigma$ is unknown but the sample size is large. However, in many applications, the sample sizes are small and we do not know the value of the standard deviation, although in some cases we might have a very good idea about its value. For example, a manufacturer of light bulbs probably has a good notion from past experience of the value of the standard deviation of the length of life of different types of light bulbs. But certainly, most of the time, the investigator will not have any more idea about the standard deviation than about the mean—and frequently less. Let us consider how to proceed under these circumstances.

If the random sample arises from a normal distribution, we use the fact that

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a $t$ distribution with $r = n - 1$ degrees of freedom (see Equation 5.5-2), where $S^2$ is the usual unbiased estimator of $\sigma^2$. Select $t_{\alpha/2}(n\text{–}1)$ so that $P[T \geq t_{\alpha/2}(n\text{–}1)] = \alpha/2$. [See Figure 5.5-2(b) and Table VI in Appendix B.] Then

$$
\begin{aligned}
1-\alpha &= P\left[-t_{\alpha/2}(n\text{–}1) \le \frac{\overline{X}-\mu}{S/\sqrt{n}} \le t_{\alpha/2}(n\text{–}1)\right] \\
&= P\left[-t_{\alpha/2}(n\text{–}1)\left(\frac{S}{\sqrt{n}}\right) \le \overline{X}-\mu \le t_{\alpha/2}(n\text{–}1)\left(\frac{S}{\sqrt{n}}\right)\right] \\
&= P\left[-\overline{X}-t_{\alpha/2}(n\text{–}1)\left(\frac{S}{\sqrt{n}}\right) \le -\mu \le -\overline{X}+t_{\alpha/2}(n\text{–}1)\left(\frac{S}{\sqrt{n}}\right)\right] \\
&= P\left[\overline{X}-t_{\alpha/2}(n\text{–}1)\left(\frac{S}{\sqrt{n}}\right) \le \mu \le \overline{X}+t_{\alpha/2}(n\text{–}1)\left(\frac{S}{\sqrt{n}}\right)\right].
\end{aligned}
$$

Thus, the observations of a random sample provide $\bar{x}$ and $s^2$, and

$$
\left[\bar{x}-t_{\alpha/2}(n\text{–}1)\left(\frac{s}{\sqrt{n}}\right), \bar{x}+t_{\alpha/2}(n\text{–}1)\left(\frac{s}{\sqrt{n}}\right)\right]
$$

is a $100(1-\alpha)\%$ confidence interval for $\mu$.

**Example 7.1-5**

Let $X$ equal the amount of butterfat in pounds produced by a typical cow during a 305-day milk production period between her first and second calves. Assume that the distribution of $X$ is $N(\mu, \sigma^2)$. To estimate $\mu$, a farmer measured the butterfat production for $n = 20$ cows and obtained the following data:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 481 | 537 | 513 | 583 | 453 | 510 | 570 | 500 | 457 | 555 |
| 618 | 327 | 350 | 643 | 499 | 421 | 505 | 637 | 599 | 392 |

For these data, $\bar{x} = 507.50$ and $s = 89.75$. Thus, a point estimate of $\mu$ is $\bar{x} = 507.50$. Since $t_{0.05}(19) = 1.729$, a 90% confidence interval for $\mu$ is

$$
507.50 \pm 1.729\left(\frac{89.75}{\sqrt{20}}\right) \quad \text{or}
$$

$$
507.50 \pm 34.70, \quad \text{or equivalently,} \quad [472.80, 542.20]. \qquad \blacksquare
$$

Let $T$ have a $t$ distribution with $n-1$ degrees of freedom. Then $t_{\alpha/2}(n\text{–}1) > z_{\alpha/2}$. Consequently, we would expect the interval $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$ to be shorter than the interval $\bar{x} \pm t_{\alpha/2}(n\text{–}1)(s/\sqrt{n})$. After all, we have more information, namely, the value of $\sigma$, in constructing the first interval. However, the length of the second interval is very much dependent on the value of $s$. If the observed $s$ is smaller than $\sigma$, a shorter confidence interval could result by the second procedure. But on the average, $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$ is the shorter of the two confidence intervals (Exercise 7.1-14).

**Example 7.1-6**

In Example 7.1-2, 50 confidence intervals were simulated for the mean of a normal distribution, assuming that the variance was known. For those same data, since $t_{0.05}(4) = 2.132$, $\bar{x} \pm 2.132(s/\sqrt{5})$ was used to calculate a 90% confidence interval for $\mu$. For those particular 50 intervals, 46 contained the mean $\mu = 50$. These 50 intervals are depicted in Figure 7.1-1(b). Note the different lengths of the intervals. Some are longer and some are shorter than the corresponding $z$ intervals. The average length of the 50 $t$ intervals is 7.137, which is quite close to the expected length of such an interval: 7.169. (See Exercise 7.1-14.) The length of the intervals that use $z$ and $\sigma = 4$ is 5.885. $\blacksquare$

If we are not able to assume that the underlying distribution is normal, but $\mu$ and $\sigma$ are both unknown, approximate confidence intervals for $\mu$ can still be constructed with the formula

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}},$$

which now only has an approximate $t$ distribution. Generally, this approximation is quite good (i.e., it is robust) for many nonnormal distributions; in particular, it works well if the underlying distribution is symmetric, unimodal, and of the continuous type. However, if the distribution is highly skewed, there is great danger in using that approximation. In such a situation, it would be safer to use certain nonparametric methods for finding a confidence interval for the median of the distribution, one of which is given in Section 7.5.

There is one other aspect of confidence intervals that should be mentioned. So far, we have created only what are called **two-sided confidence intervals** for the mean $\mu$. Sometimes, however, we might want only a lower (or upper) bound on $\mu$. We proceed as follows.

Say $\overline{X}$ is the mean of a random sample of size $n$ from the normal distribution $N(\mu, \sigma^2)$, where, for the moment, assume that $\sigma^2$ is known. Then

$$P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_\alpha\right) = 1 - \alpha,$$

or equivalently,

$$P\left[\overline{X} - z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu\right] = 1 - \alpha.$$

Once $\overline{X}$ is observed to be equal to $\bar{x}$, it follows that $[\bar{x} - z_\alpha(\sigma/\sqrt{n}), \infty)$ is a $100(1-\alpha)\%$ **one-sided confidence interval** for $\mu$. That is, with the confidence coefficient $1-\alpha$, $\bar{x} - z_\alpha(\sigma/\sqrt{n})$ is a lower bound for $\mu$. Similarly, $(-\infty, \bar{x} + z_\alpha(\sigma/\sqrt{n})]$ is a one-sided confidence interval for $\mu$ and $\bar{x} + z_\alpha(\sigma/\sqrt{n})$ provides an upper bound for $\mu$ with confidence coefficient $1-\alpha$.

When $\sigma$ is unknown, we would use $T = (\overline{X} - \mu)/(S/\sqrt{n})$ to find the corresponding lower or upper bounds for $\mu$, namely,

$$\bar{x} - t_\alpha(n-1)(s/\sqrt{n}) \quad \text{and} \quad \bar{x} + t_\alpha(n-1)(s/\sqrt{n}).$$

## Exercises

**7.1-1.** A random sample of size 16 from the normal distribution $N(\mu, 25)$ yielded $\bar{x} = 73.8$. Find a 95% confidence interval for $\mu$.

**7.1-2.** A random sample of size 8 from $N(\mu, 72)$ yielded $\bar{x} = 85$. Find the following confidence intervals for $\mu$:

**(a)** 99%. **(b)** 95%. **(c)** 90%. **(d)** 80%.

**7.1-3.** To determine the effect of 100% nitrate on the growth of pea plants, several specimens were planted and then watered with 100% nitrate every day. At the end of two weeks, the plants were measured. Here are data on seven of them:

17.5 14.5 15.2 14.0 17.3 18.0 13.8

Assume that these data are a random sample from a normal distribution $N(\mu, \sigma^2)$.

**(a)** Find the value of a point estimate of $\mu$.

**(b)** Find the value of a point estimate of $\sigma$.

**(c)** Give the endpoints for a 90% confidence interval for $\mu$.

**7.1-4.** Let $X$ equal the weight in grams of a "52-gram" snack pack of candies. Assume that the distribution of $X$ is $N(\mu, 4)$. A random sample of $n = 10$ observations of $X$ yielded the following data:

| | | | | |
|---|---|---|---|---|
| 55.95 | 56.54 | 57.58 | 55.13 | 57.48 |
| 56.06 | 59.93 | 58.30 | 52.57 | 58.46 |

**(a)** Give a point estimate for $\mu$.

**(b)** Find the endpoints for a 95% confidence interval for $\mu$.

**(c)** On the basis of these very limited data, what is the probability that an individual snack pack selected at random is filled with less than 52 grams of candy?

**7.1-5.** As a clue to the amount of organic waste in Lake Macatawa (see Example 7.1-4), a count was made of the number of bacteria colonies in 100 milliliters of water. The number of colonies, in hundreds, for $n = 30$ samples of water from the east basin yielded

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 93 | 140 | 8 | 120 | 3 | 120 | 33 | 70 | 91 | 61 |
| 7 | 100 | 19 | 98 | 110 | 23 | 14 | 94 | 57 | 9 |
| 66 | 53 | 28 | 76 | 58 | 9 | 73 | 49 | 37 | 92 |

Find an approximate 90% confidence interval for the mean number (say, $\mu_E$) of colonies in 100 milliliters of water in the east basin.

**7.1-6.** To determine whether the bacteria count was lower in the west basin of Lake Macatawa than in the east basin, $n = 37$ samples of water were taken from the west basin and the number of bacteria colonies in 100 milliliters of water was counted. The sample characteristics were $\bar{x} = 11.95$ and $s = 11.80$, measured in hundreds of colonies. Find an approximate 95% confidence interval for the mean number of colonies (say, $\mu_W$) in 100 milliliters of water in the west basin.

**7.1-7.** Thirteen tons of cheese, including "22-pound" wheels (label weight), is stored in some old gypsum mines. A random sample of $n = 9$ of these wheels yielded the following weights in pounds:

| | | | | |
|---|---|---|---|---|
| 21.50 | 18.95 | 18.55 | 19.40 | 19.15 |
| 22.35 | 22.90 | 22.20 | 23.10 | |

Assuming that the distribution of the weights of the wheels of cheese is $N(\mu, \sigma^2)$, find a 95% confidence interval for $\mu$.

**7.1-8.** Assume that the yield per acre for a particular variety of soybeans is $N(\mu, \sigma^2)$. For a random sample of $n = 5$ plots, the yields in bushels per acre were 37.4, 48.8, 46.9, 55.0, and 44.0.

**(a)** Give a point estimate for $\mu$.

**(b)** Find a 90% confidence interval for $\mu$.

**7.1-9.** During the Friday night shift, $n = 28$ mints were selected at random from a production line and weighed. They had an average weight of $\bar{x} = 21.45$ grams and a standard deviation of $s = 0.31$ grams. Give the lower endpoint of a 90% one-sided confidence interval for $\mu$, the mean weight of all the mints.

**7.1-10.** A leakage test was conducted to determine the effectiveness of a seal designed to keep the inside of a plug airtight. An air needle was inserted into the plug, and the plug and needle were placed under water. The pressure was then increased until leakage was observed. Let $X$ equal the pressure in pounds per square inch. Assume that the distribution of $X$ is $N(\mu, \sigma^2)$. The following $n = 10$ observations of $X$ were obtained:

3.1 3.3 4.5 2.8 3.5 3.5 3.7 4.2 3.9 3.3

Use the observations to

**(a)** Find a point estimate of $\mu$.

**(b)** Find a point estimate of $\sigma$.

**(c)** Find a 95% one-sided confidence interval for $\mu$ that provides an upper bound for $\mu$.

**7.1-11.** Students took $n = 35$ samples of water from the east basin of Lake Macatawa (see Example 7.1-4) and measured the amount of sodium in parts per million. For their data, they calculated $\bar{x} = 24.11$ and $s^2 = 24.44$. Find an approximate 90% confidence interval for $\mu$, the mean of the amount of sodium in parts per million.

**7.1-12.** In nuclear physics, detectors are often used to measure the energy of a particle. To calibrate a detector, particles of known energy are directed into it. The values of signals from 15 different detectors, for the same energy, are

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 260 | 216 | 259 | 206 | 265 | 284 | 291 | 229 |
| 232 | 250 | 225 | 242 | 240 | 252 | 236 | |

**(a)** Find a 95% confidence interval for $\mu$, assuming that these are observations from a $N(\mu, \sigma^2)$ distribution.

**(b)** Construct a box-and-whisker diagram of the data.

**(c)** Are these detectors doing a good job or a poor job of putting out the same signal for the same input energy?

**7.1-13.** A study was conducted to measure (1) the amount of cervical spine movement induced by different methods of gaining access to the mouth and nose to begin resuscitation of a football player who is wearing a helmet and (2) the time it takes to complete each method. One method involves using a manual screwdriver to remove the side clips holding the face mask in place and then flipping

the mask up. Twelve measured times in seconds for the manual screwdriver are

33.8 31.6 28.5 29.9 29.8 26.0 35.7 27.2 29.1 32.1 26.1 24.1

Assume that these are independent observations of a normally distributed random variable that is $N(\mu, \sigma^2)$.

**(a)** Find point estimates of $\mu$ and $\sigma$.

**(b)** Find a 95% one-sided confidence interval for $\mu$ that provides an upper bound for $\mu$.

**(c)** Does the assumption of normality seem to be justified? Why?

**7.1-14.** Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from the normal distribution $N(\mu, \sigma^2)$. Calculate the expected length of a 95% confidence interval for $\mu$, assuming that $n = 5$ and the variance is

**(a)** known.

**(b)** unknown.

HINT: To find $E(S)$, first determine $E[\sqrt{(n-1)S^2/\sigma^2}]$, recalling that $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$. (See Exercise 6.4-14.)

**7.1-15.** An automotive supplier of interior parts places several electrical wires in a harness. A pull test measures the force required to pull spliced wires apart. A customer requires that each wire spliced into the harness must withstand a pull force of 20 pounds. Let $X$ equal the pull force required to pull 20 gauge wires apart. Assume that the distribution of $X$ is $N(\mu, \sigma^2)$. The following data give 20 observations of $X$:

28.8 24.4 30.1 25.6 26.4 23.9 22.1 22.5 27.6 28.1
20.8 27.7 24.4 25.1 24.6 26.3 28.2 22.2 26.3 24.4

**(a)** Find point estimates for $\mu$ and $\sigma$.

**(b)** Find a 99% one-sided confidence interval for $\mu$ that provides a lower bound for $\mu$.

**7.1-16.** Let $S^2$ be the variance of a random sample of size $n$ from $N(\mu, \sigma^2)$. Using the fact that $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$, note that the probability

$$P\left[a \le \frac{(n-1)S^2}{\sigma^2} \le b\right] = 1 - \alpha,$$

where $a = \chi^2_{1-\alpha/2}(n-1)$ and $b = \chi^2_{\alpha/2}(n-1)$. Rewrite the inequalities to obtain

$$P\left[\frac{(n-1)S^2}{b} \le \sigma^2 \le \frac{(n-1)S^2}{a}\right] = 1 - \alpha.$$

If $n = 13$ and $12S^2 = \sum_{i=1}^{13}(x_i - \bar{x})^2 = 128.41$, show that $[6.11, 24.57]$ is a 90% **confidence interval for the variance** $\sigma^2$. Accordingly, $[2.47, 4.96]$ is a 90% confidence interval for $\sigma$.

**7.1-17.** Let $\overline{X}$ be the mean of a random sample of size $n$ from $N(\mu, 9)$. Find $n$ so that $P(\overline{X} - 1 < \mu < \overline{X} + 1) = 0.90$.

## 7.2 CONFIDENCE INTERVALS FOR THE DIFFERENCE OF TWO MEANS

Suppose that we are interested in comparing the means of two normal distributions. Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be, respectively, two independent random samples of sizes $n$ and $m$ from the two normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$. Suppose, for now, that $\sigma_X^2$ and $\sigma_Y^2$ are known. The random samples are independent; thus, the respective sample means $\overline{X}$ and $\overline{Y}$ are also independent and have distributions $N(\mu_X, \sigma_X^2/n)$ and $N(\mu_Y, \sigma_Y^2/m)$. Consequently, the distribution of $W = \overline{X} - \overline{Y}$ is $N(\mu_X - \mu_Y, \sigma_X^2/n + \sigma_Y^2/m)$ and

$$P\left(-z_{\alpha/2} \le \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \le z_{\alpha/2}\right) = 1 - \alpha,$$

which can be rewritten as

$$P[(\overline{X} - \overline{Y}) - z_{\alpha/2}\sigma_W \le \mu_X - \mu_Y \le (\overline{X} - \overline{Y}) + z_{\alpha/2}\sigma_W] = 1 - \alpha,$$

where $\sigma_W = \sqrt{\sigma_X^2/n + \sigma_Y^2/m}$ is the standard deviation of $\overline{X} - \overline{Y}$. Once the experiments have been performed and the means $\bar{x}$ and $\bar{y}$ computed, the interval

$$[\bar{x} - \bar{y} - z_{\alpha/2}\sigma_W, \bar{x} - \bar{y} + z_{\alpha/2}\sigma_W]$$

or, equivalently, $\bar{x}-\bar{y}\pm z_{\alpha/2}\sigma_W$ provides a $100(1-\alpha)\%$ confidence interval for $\mu_X-\mu_Y$. Note that this interval is centered at the point estimate $\bar{x}-\bar{y}$ of $\mu_X-\mu_Y$ and is completed by subtracting and adding the product of $z_{\alpha/2}$ and the standard deviation of the point estimator.

**Example 7.2-1** In the preceding discussion, let $n=15$, $m=8$, $\bar{x}=70.1$, $\bar{y}=75.3$, $\sigma_X^2=60$, $\sigma_Y^2=40$, and $1-\alpha=0.90$. Thus, $1-\alpha/2=0.95=\Phi(1.645)$. Hence,

$$1.645\sigma_W=1.645\sqrt{\frac{60}{15}+\frac{40}{8}}=4.935,$$

and, since $\bar{x}-\bar{y}=-5.2$, it follows that

$$[-5.2-4.935,-5.2+4.935]=[-10.135,-0.265]$$

is a 90% confidence interval for $\mu_X-\mu_Y$. Because the confidence interval does not include zero, we suspect that $\mu_Y$ is greater than $\mu_X$. ■

If the sample sizes are large and $\sigma_X$ and $\sigma_Y$ are unknown, we can replace $\sigma_X^2$ and $\sigma_Y^2$ with $s_x^2$ and $s_y^2$, where $s_x^2$ and $s_y^2$ are the values of the respective unbiased estimates of the variances. This means that

$$\bar{x}-\bar{y}\pm z_{\alpha/2}\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}$$

serves as an approximate $100(1-\alpha)\%$ confidence interval for $\mu_X-\mu_Y$.

Now consider the problem of constructing confidence intervals for the difference of the means of two normal distributions when the variances are unknown but the sample sizes are small. Let $X_1,X_2,\ldots,X_n$ and $Y_1,Y_2,\ldots,Y_m$ be two independent random samples from the distributions $N(\mu_X,\sigma_X^2)$ and $N(\mu_Y,\sigma_Y^2)$, respectively. If the sample sizes are not large (say, considerably smaller than 30), this problem can be a difficult one. However, even in these cases, if we can assume common, but unknown, variances (say, $\sigma_X^2=\sigma_Y^2=\sigma^2$), there is a way out of our difficulty.

We know that

$$Z=\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\sqrt{\sigma^2/n+\sigma^2/m}}$$

is $N(0,1)$. Moreover, since the random samples are independent,

$$U=\frac{(n-1)S_X^2}{\sigma^2}+\frac{(m-1)S_Y^2}{\sigma^2}$$

is the sum of two independent chi-square random variables; thus, the distribution of $U$ is $\chi^2(n+m-2)$. In addition, the independence of the sample means and sample variances implies that $Z$ and $U$ are independent. According to the definition of a $T$ random variable,

$$T=\frac{Z}{\sqrt{U/(n+m-2)}}$$

has a $t$ distribution with $n+m-2$ degrees of freedom. That is,

$$T = \frac{\dfrac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\sqrt{\sigma^2/n+\sigma^2/m}}}{\sqrt{\left[\dfrac{(n-1)S_X^2}{\sigma^2}+\dfrac{(m-1)S_Y^2}{\sigma^2}\right]\Big/(n+m-2)}}$$

$$= \frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\sqrt{\left[\dfrac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}\right]\left[\dfrac{1}{n}+\dfrac{1}{m}\right]}}$$

has a $t$ distribution with $r = n+m-2$ degrees of freedom. Thus, with $t_0 = t_{\alpha/2}(n+m-2)$, we have

$$P(-t_0 \le T \le t_0) = 1-\alpha.$$

Solving the inequalities for $\mu_X - \mu_Y$ yields

$$P\left(\overline{X}-\overline{Y}-t_0 S_P\sqrt{\frac{1}{n}+\frac{1}{m}} \le \mu_X-\mu_Y \le \overline{X}-\overline{Y}+t_0 S_P\sqrt{\frac{1}{n}+\frac{1}{m}}\right),$$

where the pooled estimator of the common standard deviation is

$$S_P = \sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}}.$$

If $\overline{x}$, $\overline{y}$, and $s_p$ are the observed values of $\overline{X}$, $\overline{Y}$, and $S_P$, then

$$\left[\overline{x}-\overline{y}-t_0 s_p\sqrt{\frac{1}{n}+\frac{1}{m}}, \overline{x}-\overline{y}+t_0 s_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right]$$

is a $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$.

**Example 7.2-2** Suppose that scores on a standardized test in mathematics taken by students from large and small high schools are $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$, respectively, where $\sigma^2$ is unknown. If a random sample of $n = 9$ students from large high schools yielded $\overline{x} = 81.31$, $s_x^2 = 60.76$, and a random sample of $m = 15$ students from small high schools yielded $\overline{y} = 78.61$, $s_y^2 = 48.24$, then the endpoints for a 95% confidence interval for $\mu_X - \mu_Y$ are given by

$$81.31 - 78.61 \pm 2.074\sqrt{\frac{8(60.76)+14(48.24)}{22}}\sqrt{\frac{1}{9}+\frac{1}{15}}$$

because $t_{0.025}(22) = 2.074$. The 95% confidence interval is $[-3.65, 9.05]$. ■

REMARKS The assumption of equal variances, namely, $\sigma_X^2 = \sigma_Y^2$, can be modified somewhat so that we are still able to find a confidence interval for $\mu_X - \mu_Y$. That is, if we know the ratio $\sigma_X^2/\sigma_Y^2$ of the variances, we can still make this type of statistical

inference by using a random variable with a $t$ distribution. (See Exercise 7.2-8.) However, if we do not know the ratio of the variances and yet suspect that the unknown $\sigma_X^2$ and $\sigma_Y^2$ differ by a great deal, what do we do? It is safest to return to

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}}$$

for the inference about $\mu_X - \mu_Y$ but replacing $\sigma_X^2$ and $\sigma_Y^2$ by their respective estimators $S_X^2$ and $S_Y^2$. That is, consider

$$W = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}}.$$

What is the distribution of $W$? As before, we note that if $n$ and $m$ are large enough and the underlying distributions are close to normal (or at least not badly skewed), then $W$ has an approximate normal distribution and a confidence interval for $\mu_X - \mu_Y$ can be found by considering

$$P(-z_{\alpha/2} \le W \le z_{\alpha/2}) \approx 1 - \alpha.$$

However, for smaller $n$ and $m$, Welch has proposed a Student's $t$ distribution as the approximating one for $W$. Welch's proposal was later modified by Aspin. [See A. A. Aspin, "Tables for Use in Comparisons Whose Accuracy Involves Two Variances, Separately Estimated," *Biometrika*, **36** (1949), pp. 290–296, with an appendix by B. L. Welch in which he makes the suggestion used here.] The approximating Student's $t$ distribution has $r$ degrees of freedom, where

$$\frac{1}{r} = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1} \qquad \text{and} \qquad c = \frac{s_x^2/n}{s_x^2/n + s_y^2/m}.$$

An equivalent formula for $r$ is

$$r = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{s_x^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{s_y^2}{m}\right)^2}. \tag{7.2-1}$$

In particular, the assignment of $r$ by this rule provides protection in the case in which the smaller sample size is associated with the larger variance by greatly reducing the number of degrees of freedom from the usual $n + m - 2$. Of course, this reduction increases the value of $t_{\alpha/2}$. If $r$ is not an integer, then use the greatest integer in $r$; that is, use $[r]$ as the number of degrees of freedom associated with the approximating Student's $t$ distribution. An approximate $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is given by

$$\overline{x} - \overline{y} \pm t_{\alpha/2}(r)\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}.$$

It is interesting to consider the two-sample $T$ in more detail. It is

$$T = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\dfrac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}\left(\dfrac{1}{n} + \dfrac{1}{m}\right)}} \tag{7.2-2}$$

$$= \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\left[\dfrac{(n-1)S_X^2}{nm} + \dfrac{(m-1)S_Y^2}{nm}\right]\left[\dfrac{n+m}{n+m-2}\right]}}.$$

Now, since $(n-1)/n \approx 1$, $(m-1)/m \approx 1$, and $(n+m)/(n+m-2) \approx 1$, we have

$$T \approx \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\dfrac{S_X^2}{m} + \dfrac{S_Y^2}{n}}}.$$

We note that, in this form, each variance is divided by the wrong sample size! That is, if the sample sizes are large or the variances known, we would like

$$\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} \qquad \text{or} \qquad \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

in the denominator; so $T$ seems to change the sample sizes. Thus, using this $T$ is particularly bad when the sample sizes and the variances are unequal; hence, caution must be taken in using that $T$ to construct a confidence interval for $\mu_X - \mu_Y$. That is, if $n < m$ and $\sigma_X^2 < \sigma_Y^2$, then $T$ does not have a distribution which is close to that of a Student $t$-distribution with $n + m - 2$ degrees of freedom: Instead, its spread is much less than the Student $t$'s as the term $s_y^2/n$ in the denominator is much larger than it should be. By contrast, if $m < n$ and $\sigma_X^2 < \sigma_Y^2$, then $s_x^2/m + s_y^2/n$ is generally smaller than it should be and the distribution of $T$ is spread out more than that of the Student $t$.

There is a way out of this difficulty, however: When the underlying distributions are close to normal, but the sample sizes and the variances are seemingly much different, we suggest the use of

$$W = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\dfrac{S_X^2}{n} + \dfrac{S_Y^2}{m}}}, \tag{7.2-3}$$

where Welch proved that $W$ has an approximate $t$ distribution with $[r]$ degrees of freedom, with the number of degrees of freedom given by Equation 7.2-1. ■

**Example 7.2-3**

To help understand the preceding remarks, a simulation was done with *Maple*. In order to obtain a $q$–$q$ plot of the quantiles of a $t$ distribution, a CAS or some type of computer program is very important because of the challenge in finding these quantiles.

*Maple* was used to simulate $N = 500$ observations of $T$ (Equation 7.2-2) and $N = 500$ observations of $W$ (Equation 7.2-3). In Figure 7.2-1, $n = 6$, $m = 18$, the

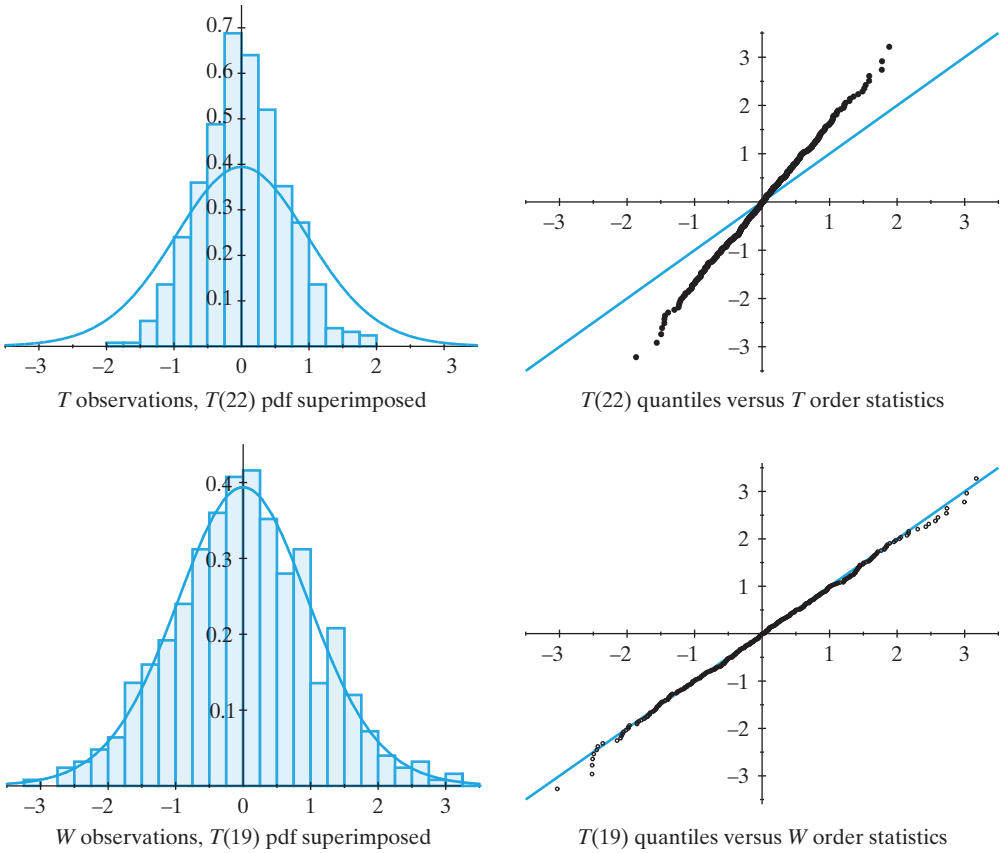



**Figure 7.2-1** Observations of $T$ and of $W$, $n = 6$, $m = 18$, $\sigma_X^2 = 1$, $\sigma_Y^2 = 36$

$X$ observations were generated from the $N(0,1)$ distribution, and the $Y$ observations were generated from the $N(0,36)$ distribution. For the value of $r$ for Welch's approximate $t$ distribution, we used the distribution variances rather than the sample variances so that we could use the same $r$ for each of the 500 values of $W$.

For the simulation results shown in Figure 7.2-2, $n = 18$, $m = 6$, the $X$ observations were generated from the $N(0,1)$ distribution, and the $Y$ observations were generated from the $N(0,36)$ distribution. In both cases, Welch's $W$ with a corrected number of $r$ degrees of freedom is much better than the usual $T$ when the variances and sample sizes are unequal, as they are in these examples. ■

In some applications, two measurements—say, $X$ and $Y$—are taken on the same subject. In these cases, $X$ and $Y$ may be dependent random variables. Many times these are "before" and "after" measurements, such as weight before and after participating in a diet-and-exercise program. To compare the means of $X$ and $Y$, it is not permissible to use the $t$ statistics and confidence intervals that we just developed, because in that situation $X$ and $Y$ are independent. Instead, we proceed as follows.

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be $n$ pairs of dependent measurements. Let $D_i = X_i - Y_i$, $i = 1, 2, \ldots, n$. Suppose that $D_1, D_2, \ldots, D_n$ can be thought of as

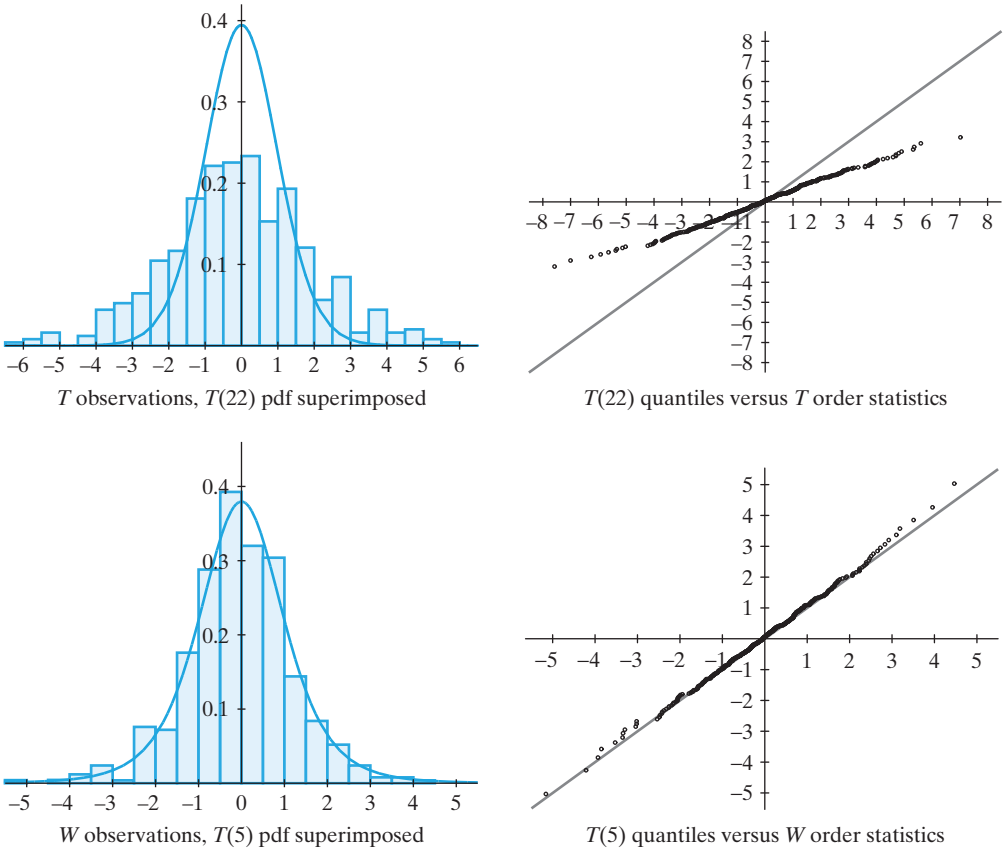



**Figure 7.2-2** Observations of $T$ and of $W$, $n = 18$, $m = 6$, $\sigma_X^2 = 1$, $\sigma_Y^2 = 36$

a random sample from $N(\mu_D, \sigma_D^2)$, where $\mu_D$ and $\sigma_D$ are the mean and standard deviation of each difference. To form a confidence interval for $\mu_X - \mu_Y$, use

$$T = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}},$$

where $\overline{D}$ and $S_D$ are, respectively, the sample mean and sample standard deviation of the $n$ differences. Thus, $T$ is a $t$ statistic with $n-1$ degrees of freedom. The endpoints for a $100(1-\alpha)\%$ confidence interval for $\mu_D = \mu_X - \mu_Y$ are then

$$\overline{d} \pm t_{\alpha/2}(n-1)\,\frac{s_d}{\sqrt{n}},$$

where $\overline{d}$ and $s_d$ are the observed mean and standard deviation of the sample of the $D$ values. Of course, this is like the confidence interval for a single mean, presented in the last section.

**Example 7.2-4** An experiment was conducted to compare people's reaction times to a red light versus a green light. When signaled with either the red or the green light, the subject was asked to hit a switch to turn off the light. When the switch was hit, a clock was turned off and the reaction time in seconds was recorded. The following results give the reaction times for eight subjects:

| Subject | Red ($x$) | Green ($y$) | $d = x - y$ |
|---|---|---|---|
| 1 | 0.30 | 0.43 | –0.13 |
| 2 | 0.23 | 0.32 | –0.09 |
| 3 | 0.41 | 0.58 | –0.17 |
| 4 | 0.53 | 0.46 | 0.07 |
| 5 | 0.24 | 0.27 | –0.03 |
| 6 | 0.36 | 0.41 | –0.05 |
| 7 | 0.38 | 0.38 | 0.00 |
| 8 | 0.51 | 0.61 | –0.10 |

For these data, $\overline{d} = -0.0625$ and $s_d = 0.0765$. To form a 95% confidence interval for $\mu_D = \mu_X - \mu_Y$, we find, from Table VI in Appendix B, that $t_{0.025}(7) = 2.365$. Thus, the endpoints for the confidence interval are

$$-0.0625 \pm 2.365 \frac{0.0765}{\sqrt{8}}, \qquad \text{or} \qquad [-0.1265, 0.0015].$$

In this very limited data set, zero is included in the confidence interval but is close to the endpoint 0.0015. We suspect that if more data were taken, zero might not be included in the confidence interval. If that actually were to happen, it would seem that people react faster to a red light. ■

Of course, we can find one-sided confidence intervals for the difference of the means, $\mu_X - \mu_Y$. Suppose we believe that we have changed some characteristic of the $X$ distribution and created a $Y$ distribution such that we think that $\mu_X > \mu_Y$. Let us find a one-sided 95% confidence interval that is a lower bound for $\mu_X - \mu_Y$. Say this lower bound is greater than zero. Then we would feel 95% confident that the mean $\mu_X$ is larger than the mean $\mu_Y$. That is, the change that was made seemed to decrease the mean; this would be good in some cases, such as golf or racing. In other cases, in which we hope the change would be such that $\mu_X < \mu_Y$, we would find a one-sided confidence interval which is an upper bound for $\mu_X - \mu_Y$, and we would hope that it would be less than zero. These ideas are illustrated in Exercises 7.2-5, 7.2-10, and 7.2-11.

## Exercises

**7.2-1.** The length of life of brand $X$ light bulbs is assumed to be $N(\mu_X, 784)$. The length of life of brand $Y$ light bulbs is assumed to be $N(\mu_Y, 627)$ and independent of $X$. If a random sample of $n = 56$ brand $X$ light bulbs yielded a mean of $\bar{x} = 937.4$ hours and a random sample of size $m = 57$ brand $Y$ light bulbs yielded a mean of $\bar{y} = 988.9$ hours, find a 90% confidence interval for $\mu_X - \mu_Y$.

**7.2-2.** Let $X_1, X_2, \ldots, X_5$ be a random sample of SAT mathematics scores, assumed to be $N(\mu_X, \sigma^2)$, and let $Y_1, Y_2, \ldots, Y_8$ be an independent random sample of SAT

verbal scores, assumed to be $N(\mu_Y, \sigma^2)$. If the following data are observed, find a 90% confidence interval for $\mu_X - \mu_Y$:

| | | | | |
|---|---|---|---|---|
| $x_1 = 644$ | $x_2 = 493$ | $x_3 = 532$ | $x_4 = 462$ | $x_5 = 565$ |
| $y_1 = 623$ | $y_2 = 472$ | $y_3 = 492$ | $y_4 = 661$ | $y_5 = 540$ |
| $y_6 = 502$ | $y_7 = 549$ | $y_8 = 518$ | | |

**7.2-3.** Independent random samples of the heights of adult males living in two countries yielded the following results: $n = 12$, $\bar{x} = 65.7$ inches, $s_x = 4$ inches and $m = 15$, $\bar{y} = 68.2$ inches, $s_y = 3$ inches. Find an approximate 98% confidence interval for the difference $\mu_X - \mu_Y$ of the means of the populations of heights. Assume that $\sigma_X^2 = \sigma_Y^2$.

**7.2-4.** [*Medicine and Science in Sports and Exercise* (January 1990).] Let $X$ and $Y$ equal, respectively, the blood volumes in milliliters for a male who is a paraplegic and participates in vigorous physical activities and for a male who is able-bodied and participates in everyday, ordinary activities. Assume that $X$ is $N(\mu_X, \sigma_X^2)$ and $Y$ is $N(\mu_Y, \sigma_Y^2)$. Following are $n = 7$ observations of $X$:

| | | | | | | |
|---|---|---|---|---|---|---|
| 1612 | 1352 | 1456 | 1222 | 1560 | 1456 | 1924 |

Following are $m = 10$ observations of $Y$:

| | | | | |
|---|---|---|---|---|
| 1082 | 1300 | 1092 | 1040 | 910 |
| 1248 | 1092 | 1040 | 1092 | 1288 |

Use the observations of $X$ and $Y$ to

**(a)** Give a point estimate for $\mu_X - \mu_Y$.

**(b)** Find a 95% confidence interval for $\mu_X - \mu_Y$. Since the variances $\sigma_X^2$ and $\sigma_Y^2$ might not be equal, use Welch's $T$.

**7.2-5.** A biologist who studies spiders was interested in comparing the lengths of female and male green lynx spiders. Assume that the length $X$ of the male spider is approximately $N(\mu_X, \sigma_X^2)$ and the length $Y$ of the female spider is approximately $N(\mu_Y, \sigma_Y^2)$. Following are $n = 30$ observations of $X$:

| | | | | | |
|---|---|---|---|---|---|
| 5.20 | 4.70 | 5.75 | 7.50 | 6.45 | 6.55 |
| 4.70 | 4.80 | 5.95 | 5.20 | 6.35 | 6.95 |
| 5.70 | 6.20 | 5.40 | 6.20 | 5.85 | 6.80 |
| 5.65 | 5.50 | 5.65 | 5.85 | 5.75 | 6.35 |
| 5.75 | 5.95 | 5.90 | 7.00 | 6.10 | 5.80 |

Following are $m = 30$ observations of $Y$:

| | | | | | |
|---|---|---|---|---|---|
| 8.25 | 9.95 | 5.90 | 7.05 | 8.45 | 7.55 |
| 9.80 | 10.80 | 6.60 | 7.55 | 8.10 | 9.10 |
| 6.10 | 9.30 | 8.75 | 7.00 | 7.80 | 8.00 |
| 9.00 | 6.30 | 8.35 | 8.70 | 8.00 | 7.50 |
| 9.50 | 8.30 | 7.05 | 8.30 | 7.95 | 9.60 |

The units of measurement for both sets of observations are millimeters. Find an approximate one-sided 95% confidence interval that is an upper bound for $\mu_X - \mu_Y$.

**7.2-6.** A test was conducted to determine whether a wedge on the end of a plug fitting designed to hold a seal onto the plug was doing its job. The data taken were in the form of measurements of the force required to remove a seal from the plug with the wedge in place (say, $X$) and the force required without the plug (say, $Y$). Assume that the distributions of $X$ and $Y$ are $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$, respectively. Ten independent observations of $X$ are

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 3.26 | 2.26 | 2.62 | 2.62 | 2.36 | 3.00 | 2.62 | 2.40 | 2.30 | 2.40 |

Ten independent observations of $Y$ are

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1.80 | 1.46 | 1.54 | 1.42 | 1.32 | 1.56 | 1.36 | 1.64 | 2.00 | 1.54 |

**(a)** Find a 95% confidence interval for $\mu_X - \mu_Y$.

**(b)** Construct box-and-whisker diagrams of these data on the same figure.

**(c)** Is the wedge necessary?

**7.2-7.** An automotive supplier is considering changing its electrical wire harness to save money. The idea is to replace a current 20-gauge wire with a 22-gauge wire. Since not all wires in the harness can be changed, the new wire must work with the current wire splice process. To determine whether the new wire is compatible, random samples were selected and measured with a pull test. A pull test measures the force required to pull the spliced wires apart. The minimum pull force required by the customer is 20 pounds. Twenty observations of the forces needed for the current wire are

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 28.8 | 24.4 | 30.1 | 25.6 | 26.4 | 23.9 | 22.1 | 22.5 | 27.6 | 28.1 |
| 20.8 | 27.7 | 24.4 | 25.1 | 24.6 | 26.3 | 28.2 | 22.2 | 26.3 | 24.4 |

Twenty observations of the forces needed for the new wire are

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 14.1 | 12.2 | 14.0 | 14.6 | 8.5 | 12.6 | 13.7 | 14.8 | 14.1 | 13.2 |
| 12.1 | 11.4 | 10.1 | 14.2 | 13.6 | 13.1 | 11.9 | 14.8 | 11.1 | 13.5 |

**(a)** Does the current wire meet the customer's specifications?

**(b)** Find a 90% confidence interval for the difference of the means for these two sets of wire.

**(c)** Construct box-and-whisker diagrams of the two sets of data on the same figure.

**(d)** What is your recommendation for this company?

**7.2-8.** Let $\overline{X}$, $\overline{Y}$, $S_X^2$, and $S_Y^2$ be the respective sample means and unbiased estimates of the variances obtained from independent samples of sizes $n$ and $m$ from the normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, where $\mu_X$, $\mu_Y$, $\sigma_X^2$, and $\sigma_Y^2$ are unknown. If $\sigma_X^2/\sigma_Y^2 = d$, a known constant,

**(a)** Argue that $\dfrac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sqrt{d\sigma_Y^2/n+\sigma_Y^2/m}}$ is $N(0,1)$.

**(b)** Argue that $\dfrac{(n-1)S_X^2}{d\sigma_Y^2}+\dfrac{(m-1)S_Y^2}{\sigma_Y^2}$ is $\chi^2(n+m-2)$.

**(c)** Argue that the two random variables in (a) and (b) are independent.

**(d)** With these results, construct a random variable (not depending upon $\sigma_Y^2$) that has a $t$ distribution and that can be used to construct a confidence interval for $\mu_X - \mu_Y$.

**7.2-9.** Students in a semester-long health-fitness program have their percentage of body fat measured at the beginning of the semester and at the end of the semester. The following measurements give these percentages for 10 men and for 10 women:

| Males | | Females | |
|---|---|---|---|
| Pre-program % | Post-program % | Pre-program % | Post-program % |
| 11.10 | 9.97 | 22.90 | 22.89 |
| 19.50 | 15.80 | 31.60 | 33.47 |
| 14.00 | 13.02 | 27.70 | 25.75 |
| 8.30 | 9.28 | 21.70 | 19.80 |
| 12.40 | 11.51 | 19.36 | 18.00 |
| 7.89 | 7.40 | 25.03 | 22.33 |
| 12.10 | 10.70 | 26.90 | 25.26 |
| 8.30 | 10.40 | 25.75 | 24.90 |
| 12.31 | 11.40 | 23.63 | 21.80 |
| 10.00 | 11.95 | 25.06 | 24.28 |

**(a)** Find a 90% confidence interval for the mean of the difference in the percentages for the males.

**(b)** Find a 90% confidence interval for the mean of the difference in the percentages for the females.

**(c)** On the basis of these data, have these percentages decreased?

**(d)** If possible, check whether each set of differences comes from a normal distribution.

**7.2-10.** Twenty-four 9th- and 10th-grade high school girls were put on an ultraheavy rope-jumping program. The following data give the time difference for each girl ("before program time" minus "after program time") for the 40-yard dash:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0.28 | 0.01 | 0.13 | 0.33 | –0.03 | 0.07 | –0.18 | –0.14 |
| –0.33 | 0.01 | 0.22 | 0.29 | –0.08 | 0.23 | 0.08 | 0.04 |
| –0.30 | –0.08 | 0.09 | 0.70 | 0.33 | –0.34 | 0.50 | 0.06 |

**(a)** Give a point estimate of $\mu_D$, the mean of the difference in race times.

**(b)** Find a one-sided 95% confidence interval that is a lower bound for $\mu_D$.

**(c)** Does it look like the rope-jumping program was effective?

**7.2-11.** The Biomechanics Lab at Hope College tested healthy old women and healthy young women to discover whether or not lower extremity response time to a stimulus is a function of age. Let $X$ and $Y$ respectively equal the independent response times for these two groups when taking steps in the anterior direction. Find a one-sided 95% confidence interval that is a lower bound for $\mu_X - \mu_Y$ if $n = 60$ observations of $X$ yielded $\bar{x} = 671$ and $s_x = 129$, while $m = 60$ observations of $Y$ yielded $\bar{y} = 480$ and $s_y = 93$.

**7.2-12.** Let $X$ and $Y$ equal the hardness of the hot and cold water, respectively, in a campus building. Hardness is measured in terms of the calcium ion concentration (in ppm). The following data were collected ($n = 12$ observations of $X$ and $m = 10$ observations of $Y$):

| | | | | | | |
|---|---|---|---|---|---|---|
| $x$: | 133.5 | 137.2 | 136.3 | 133.3 | 137.5 | 135.4 |
| | 138.4 | 137.1 | 136.5 | 139.4 | 137.9 | 136.8 |
| $y$: | 134.0 | 134.7 | 136.0 | 132.7 | 134.6 | 135.2 |
| | 135.9 | 135.6 | 135.8 | 134.2 | | |

**(a)** Calculate the sample means and the sample variances of these data.

**(b)** Construct a 95% confidence interval for $\mu_X - \mu_Y$, assuming that the distributions of $X$ and $Y$ are $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, respectively.

**(c)** Construct box plots of the two sets of data on the same graph.

**(d)** Do the means seem to be equal or different?

**7.2-13.** Ledolter and Hogg (see References) report that two rubber compounds were tested for tensile strength. Rectangular materials were prepared and pulled in a longitudinal direction. A sample of 14 specimens, 7 from compound $A$ and 7 from compound $B$, was prepared, but it was later found that two $B$ specimens were defective and they had to be removed from the test. The tensile strength (in units of 100 pounds per square inch) of the remaining specimens are as follows:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $A$: | 32 | 30 | 33 | 32 | 29 | 34 | 32 |
| $B$: | 33 | 35 | 36 | 37 | 35 | | |

Calculate a 95% confidence interval for the difference of the mean tensile strengths of the two rubber compounds. State your assumptions.

**7.2-14.** Let $S_X^2$ and $S_Y^2$ be the respective variances of two independent random samples of sizes $n$ and $m$ from $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$. Use the fact that $F = [S_Y^2/\sigma_Y^2]/[S_X^2/\sigma_X^2]$ has an $F$ distribution, with parameters $r_1 = m-1$ and $r_2 = n-1$, to rewrite $P(c \le F \le d) = 1-\alpha$, where $c = F_{1-\alpha/2}(r_1, r_2)$ and $d = F_{\alpha/2}(r_1, r_2)$, so that

$$P\left(c\frac{S_X^2}{S_Y^2} \le \frac{\sigma_X^2}{\sigma_Y^2} \le d\frac{S_X^2}{S_Y^2}\right) = 1-\alpha.$$

If the observed values are $n = 13$, $m = 9$, $12s_x^2 = 128.41$, and $8s_y^2 = 36.72$, show that a 98% **confidence interval for the ratio of the two variances**, $\sigma_X^2/\sigma_Y^2$, is $[0.41, 10.49]$, so that $[0.64, 3.24]$ is a 98% confidence interval for $\sigma_X/\sigma_Y$.

## 7.3 CONFIDENCE INTERVALS FOR PROPORTIONS

We have suggested that the histogram is a good description of how the observations of a random sample are distributed. We might naturally inquire about the accuracy of those relative frequencies (or percentages) associated with the various classes. To illustrate, in Example 6.1-1 concerning the weights of $n = 40$ candy bars, we found that the relative frequency of the class interval (22.25, 23.15) was $8/40 = 0.20$, or 20%. If we think of this collection of 40 weights as a random sample observed from a larger population of candy bar weights, how close is 20% to the true percentage (or 0.20 to the true proportion) of weights in that class interval for the entire population of weights for this type of candy bar?

In considering this problem, we generalize it somewhat by treating the class interval (22.25, 23.15) as "success." That is, there is some true probability of success, $p$—namely, the proportion of the population in that interval. Let $Y$ equal the frequency of measurements in the interval out of the $n$ observations, so that (under the assumptions of independence and constant probability $p$) $Y$ has the binomial distribution $b(n, p)$. Thus, the problem is to determine the accuracy of the relative frequency $Y/n$ as an estimator of $p$. We solve this problem by finding, for the unknown $p$, a confidence interval based on $Y/n$.

In general, when observing $n$ Bernoulli trials with probability $p$ of success on each trial, we shall find a confidence interval for $p$ based on $Y/n$, where $Y$ is the number of successes and $Y/n$ is an unbiased point estimator for $p$.

In Section 5.7, we noted that

$$\frac{Y - np}{\sqrt{np(1-p)}} = \frac{(Y/n) - p}{\sqrt{p(1-p)/n}}$$

has an approximate normal distribution $N(0, 1)$, provided that $n$ is large enough. This means that, for a given probability $1-\alpha$, we can find a $z_{\alpha/2}$ in Table V in Appendix B such that

$$P\left[-z_{\alpha/2} \le \frac{(Y/n)-p}{\sqrt{p(1-p)/n}} \le z_{\alpha/2}\right] \approx 1-\alpha. \tag{7.3-1}$$

If we proceed as we did when we found a confidence interval for $\mu$ in Section 7.1, we would obtain

$$P\left[\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p \le \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right] \approx 1-\alpha.$$

Unfortunately, the unknown parameter $p$ appears in the endpoints of this inequality. There are two ways out of this dilemma. First, we could make an additional approximation, namely, replacing $p$ with $Y/n$ in $p\,(1-p)/n$ in the endpoints. That is, if $n$ is large enough, it is still true that

$$P\left[\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{(Y/n)(1-Y/n)}{n}} \le p \le \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{(Y/n)(1-Y/n)}{n}}\right] \approx 1-\alpha.$$

Thus, for large $n$, if the observed $Y$ equals $y$, then the interval

$$\left[\frac{y}{n} - z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}, \frac{y}{n} + z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}\right]$$

serves as an approximate $100(1-\alpha)\%$ confidence interval for $p$. Frequently, this interval is written as

$$\frac{y}{n} \pm z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}} \tag{7.3-2}$$

for brevity. This formulation clearly notes, as does $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$ in Section 7.1, the reliability of the estimate $y/n$, namely, that we are $100(1-\alpha)\%$ confident that $p$ is within $z_{\alpha/2}\sqrt{(y/n)(1-y/n)/n}$ of $\widehat{p} = y/n$.

A second way to solve for $p$ in the inequality in Equation 7.3-1 is to note that

$$\frac{|Y/n - p|}{\sqrt{p\,(1-p)/n}} \le z_{\alpha/2}$$

is equivalent to

$$H(p) = \left(\frac{Y}{n} - p\right)^2 - \frac{z_{\alpha/2}^2\, p(1-p)}{n} \le 0. \tag{7.3-3}$$

But $H(p)$ is a quadratic expression in $p$. Thus, we can find those values of $p$ for which $H(p) \le 0$ by finding the two zeros of $H(p)$. Letting $\widehat{p} = Y/n$ and $z_0 = z_{\alpha/2}$ in Equation 7.3-3, we have

$$H(p) = \left(1 + \frac{z_0^2}{n}\right)p^2 - \left(2\widehat{p} + \frac{z_0^2}{n}\right)p + \widehat{p}^{\,2}.$$

By the quadratic formula, the zeros of $H(p)$ are, after simplifications,

$$\frac{\widehat{p} + z_0^2/(2n) \pm z_0\sqrt{\widehat{p}\,(1-\widehat{p}\,)/n + z_0^2/(4n^2)}}{1 + z_0^2/n}, \tag{7.3-4}$$

and these zeros give the endpoints for an approximate $100(1-\alpha)\%$ confidence interval for $p$. If $n$ is large, $z_0^2/(2n)$, $z_0^2/(4n^2)$, and $z_0^2/n$ are small. Thus, the confidence intervals given by Equations 7.3-2 and 7.3-4 are approximately equal when $n$ is large.

**Example 7.3-1**

Let us return to the example of the histogram of the candy bar weights, Example 6.1-1, with $n = 40$ and $y/n = 8/40 = 0.20$. If $1-\alpha = 0.90$, so that $z_{\alpha/2} = 1.645$, then, using Equation 7.3-2, we find that the endpoints

$$0.20 \pm 1.645\sqrt{\frac{(0.20)(0.80)}{40}}$$

serve as an approximate 90% confidence interval for the true fraction $p$. That is, $[0.096, 0.304]$, which is the same as $[9.6\%, 30.4\%]$, is an approximate 90% confidence interval for the percentage of weights of the entire population in the interval $(22.25, 23.15)$. If we had used the endpoints given by Equation 7.3-4, the confidence interval would be $[0.117, 0.321]$. Because of the small sample size, there is a non-negligible difference in these intervals. If the sample size had been $n = 400$ and $y = 80$, so that $y/n = 80/400 = 0.20$, the two 90% confidence intervals would have been $[0.167, 0.233]$ and $[0.169, 0.235]$, respectively, which differ very little. ■

**Example 7.3-2**

In a certain political campaign, one candidate has a poll taken at random among the voting population. The results are that $y = 185$ out of $n = 351$ voters favor this candidate. Even though $y/n = 185/351 = 0.527$, should the candidate feel very confident of winning? From Equation 7.3-2, an approximate 95% confidence interval for the fraction $p$ of the voting population who favor the candidate is

$$0.527 \pm 1.96\sqrt{\frac{(0.527)(0.473)}{351}}$$

or, equivalently, $[0.475, 0.579]$. Thus, there is a good possibility that $p$ is less than 50%, and the candidate should certainly take this possibility into account in campaigning. ■

One-sided confidence intervals are sometimes appropriate for $p$. For example, we may be interested in an upper bound on the proportion of defectives in manufacturing some item. Or we may be interested in a lower bound on the proportion of voters who favor a particular candidate. The one-sided confidence interval for $p$ given by

$$\left[0, \frac{y}{n} + z_\alpha\sqrt{\frac{(y/n)[1-(y/n)]}{n}}\right]$$

provides an upper bound for $p$, while

$$\left[\frac{y}{n} - z_\alpha\sqrt{\frac{(y/n)[1-(y/n)]}{n}}, 1\right]$$

provides a lower bound for $p$.

REMARK Sometimes the confidence intervals suggested here are not very close to having the stated confidence coefficient. This is particularly true if $n$ is small or if one of $Y$ or $n - Y$ is close to zero. It is obvious that something is wrong if $Y = 0$ or $n - Y = 0$, because the radical is then equal to zero.

It has been suggested (see, e.g., Agresti and Coull, 1998) that we use $\tilde{p} = (Y + 2)/(n + 4)$ as an estimator for $p$ in those cases because the results are usually much better. It is true that $\tilde{p}$ is a biased estimator of $p$, but it is a Bayes shrinkage estimator if we use the beta prior pdf with parameters $\alpha = 2$, $\beta = 2$. In those cases in which $n$ is small or $Y$ or $n - Y$ is close to zero,

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/(n+4)} \tag{7.3-5}$$

provides a much better $100(1 - \alpha)\%$ confidence interval for $p$. A similar statement can be made about one-sided confidence intervals.

Look again at Equation 7.3-4. If we form a 95% confidence interval using this equation, we find that $z_0 = 1.96 \approx 2$. Thus, a 95% confidence interval is centered approximately at

$$\frac{\hat{p} + z_0^2/(2n)}{1 + z_0^2/n} = \frac{y + z_0^2/2}{n + z_0^2} \approx \frac{y+2}{n+4}.$$

This result is consistent with Equation 7.3-5 for 95% confidence intervals. ■

**Example 7.3-3** Returning to the data in Example 7.3-1, and using Equation 7.3-5, we have $\tilde{p} = (8 + 2)/(40 + 4) = 0.227$. Thus, a 90% confidence interval is

$$0.227 \pm 1.645\sqrt{\frac{(0.227)(0.773)}{44}},$$

or $[0.123, 0.331]$. If it had been true that $y = 80$ and $n = 400$, the confidence interval given by Equation 7.3-5 would have been $[0.170, 0.236]$. ■

Frequently, there are two (or more) possible independent ways of performing an experiment; suppose these have probabilities of success $p_1$ and $p_2$, respectively. Let $n_1$ and $n_2$ be the number of independent trials associated with these two methods, and let us say that they result in $Y_1$ and $Y_2$ successes, respectively. In order to make a statistical inference about the difference $p_1 - p_2$, we proceed as follows.

Since the independent random variables $Y_1/n_1$ and $Y_2/n_2$ have respective means $p_1$ and $p_2$ and variances $p_1(1 - p_1)/n_1$ and $p_2(1 - p_2)/n_2$, we know from Section 5.4 that the difference $Y_1/n_1 - Y_2/n_2$ must have mean $p_1 - p_2$ and variance

$$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

(Recall that the variances are added to get the variance of a difference of two independent random variables.) Moreover, the fact that $Y_1/n_1$ and $Y_2/n_2$ have approximate normal distributions would suggest that the difference

$$\frac{Y_1}{n_1} - \frac{Y_2}{n_2}$$

would have an approximate normal distribution with the above mean and variance. (See Theorem 5.5-1.) That is,

$$\frac{(Y_1/n_1)-(Y_2/n_2)-(p_1-p_2)}{\sqrt{p_1(1-p_1)/n_1+p_2(1-p_2)/n_2}}$$

has an approximate normal distribution $N(0,1)$. If we now replace $p_1$ and $p_2$ in the denominator of this ratio by $Y_1/n_1$ and $Y_2/n_2$, respectively, it is still true for large enough $n_1$ and $n_2$ that the new ratio will be approximately $N(0,1)$. Thus, for a given $1-\alpha$, we can find $z_{\alpha/2}$ from Table V in Appendix B, so that

$$P\left[-z_{\alpha/2}\le\frac{(Y_1/n_1)-(Y_2/n_2)-(p_1-p_2)}{\sqrt{(Y_1/n_1)(1-Y_1/n_1)/n_1+(Y_2/n_2)(1-Y_2/n_2)/n_2}}\le z_{\alpha/2}\right]\approx 1-\alpha.$$

Once $Y_1$ and $Y_2$ are observed to be $y_1$ and $y_2$, respectively, this approximation can be solved to obtain an approximate $100(1-\alpha)\%$ confidence interval

$$\frac{y_1}{n_1}-\frac{y_2}{n_2}\pm z_{\alpha/2}\sqrt{\frac{(y_1/n_1)(1-y_1/n_1)}{n_1}+\frac{(y_2/n_2)(1-y_2/n_2)}{n_2}}$$

for the unknown difference $p_1-p_2$. Note again how this form indicates the reliability of the estimate $y_1/n_1-y_2/n_2$ of the difference $p_1-p_2$.

**Example 7.3-4**

Two detergents were tested for their ability to remove stains of a certain type. An inspector judged the first one to be successful on 63 out of 91 independent trials and the second one to be successful on 42 out of 79 independent trials. The respective relative frequencies of success are 63/91 = 0.692 and 42/79 = 0.532. An approximate 90% confidence interval for the difference $p_1-p_2$ of the two detergents is

$$\left(\frac{63}{91}-\frac{42}{79}\right)\pm 1.645\sqrt{\frac{(63/91)(28/91)}{91}+\frac{(42/79)(37/79)}{79}}$$

or, equivalently, $[0.039, 0.283]$. Accordingly, since this interval does not include zero, it seems that the first detergent is probably better than the second one for removing the type of stains in question. ■

## Exercises

**7.3-1.** A machine shop manufactures toggle levers. A lever is flawed if a standard nut cannot be screwed onto the threads. Let $p$ equal the proportion of flawed toggle levers that the shop manufactures. If there were 24 flawed levers out of a sample of 642 that were selected randomly from the production line,

**(a)** Give a point estimate of $p$.

**(b)** Use Equation 7.3-2 to find an approximate 95% confidence interval for $p$.

**(c)** Use Equation 7.3-4 to find an approximate 95% confidence interval for $p$.

**(d)** Use Equation 7.3-5 to find an approximate 95% confidence interval for $p$.

**(e)** Find a one-sided 95% confidence interval for $p$ that provides an upper bound for $p$.

**7.3-2.** Let $p$ equal the proportion of letters mailed in the Netherlands that are delivered the next day. Suppose that

$y = 142$ out of a random sample of $n = 200$ letters were delivered the day after they were mailed.

**(a)** Give a point estimate of $p$.

**(b)** Use Equation 7.3-2 to find an approximate 90% confidence interval for $p$.

**(c)** Use Equation 7.3-4 to find an approximate 90% confidence interval for $p$.

**(d)** Use Equation 7.3-5 to find an approximate 90% confidence interval for $p$.

**(e)** Find a one-sided 90% confidence interval for $p$ that provides a lower bound for $p$.

**7.3-3.** Let $p$ equal the proportion of triathletes who suffered a training-related overuse injury during the past year. Out of 330 triathletes who responded to a survey, 167 indicated that they had suffered such an injury during the past year.

**(a)** Use these data to give a point estimate of $p$.

**(b)** Use these data to find an approximate 90% confidence interval for $p$.

**(c)** Do you think that the 330 triathletes who responded to the survey may be considered a random sample from the population of triathletes?

**7.3-4.** Let $p$ equal the proportion of Americans who favor the death penalty. If a random sample of $n = 1234$ Americans yielded $y = 864$ who favored the death penalty, find an approximate 95% confidence interval for $p$.

**7.3-5.** In order to estimate the proportion, $p$, of a large class of college freshmen that had high school GPAs from 3.2 to 3.6, inclusive, a sample of $n = 50$ students was taken. It was found that $y = 9$ students fell into this interval.

**(a)** Give a point estimate of $p$.

**(b)** Use Equation 7.3-2 to find an approximate 95% confidence interval for $p$.

**(c)** Use Equation 7.3-4 to find an approximate 95% confidence interval for $p$.

**(d)** Use Equation 7.3-5 to find an approximate 95% confidence interval for $p$.

**7.3-6.** Let $p$ equal the proportion of Americans who select jogging as one of their recreational activities. If 1497 out of a random sample of 5757 selected jogging, find an approximate 98% confidence interval for $p$.

**7.3-7.** In developing countries in Africa and the Americas, let $p_1$ and $p_2$ be the respective proportions of women with nutritional anemia. Find an approximate 90% confidence interval for $p_1 - p_2$, given that a random sample of $n_1 = 2100$ African women yielded $y_1 = 840$ with nutritional anemia and a random sample of $n_2 = 1900$ women from the Americas yielded $y_2 = 323$ women with nutritional anemia.

**7.3-8.** A proportion, $p$, that many public opinion polls estimate is the number of Americans who would say yes to the question, "If something were to happen to the president of the United States, do you think that the vice president would be qualified to take over as president?" In one such random sample of 1022 adults, 388 said yes.

**(a)** On the basis of the given data, find a point estimate of $p$.

**(b)** Find an approximate 90% confidence interval for $p$.

**(c)** Give updated answers to this question if new poll results are available.

**7.3-9.** Consider the following two groups of women: Group 1 consists of women who spend less than \$500 annually on clothes; Group 2 comprises women who spend over \$1000 annually on clothes. Let $p_1$ and $p_2$ equal the proportions of women in these two groups, respectively, who believe that clothes are too expensive. If 1009 out of a random sample of 1230 women from group 1 and 207 out of a random sample 340 from group 2 believe that clothes are too expensive,

**(a)** Give a point estimate of $p_1 - p_2$.

**(b)** Find an approximate 95% confidence interval for $p_1 - p_2$.

**7.3-10.** A candy manufacturer selects mints at random from the production line and weighs them. For one week, the day shift weighed $n_1 = 194$ mints and the night shift weighed $n_2 = 162$ mints. The numbers of these mints that weighed at most 21 grams was $y_1 = 28$ for the day shift and $y_2 = 11$ for the night shift. Let $p_1$ and $p_2$ denote the proportions of mints that weigh at most 21 grams for the day and night shifts, respectively.

**(a)** Give a point estimate of $p_1$.

**(b)** Give the endpoints for a 95% confidence interval for $p_1$.

**(c)** Give a point estimate of $p_1 - p_2$.

**(d)** Find a one-sided 95% confidence interval that gives a lower bound for $p_1 - p_2$.

**7.3-11.** For developing countries in Asia (excluding China) and Africa, let $p_1$ and $p_2$ be the respective proportions of preschool children with chronic malnutrition (stunting). If respective random samples of $n_1 = 1300$ and $n_2 = 1100$ yielded $y_1 = 520$ and $y_2 = 385$ children with chronic malnutrition, find an approximate 95% confidence interval for $p_1 - p_2$.

**7.3-12.** An environmental survey contained a question asking what respondents thought was the major cause of air pollution in this country, giving the choices "automobiles," "factories," and "incinerators." Two versions of the test, $A$ and $B$, were used. Let $p_A$ and $p_B$ be the respective proportions of people using forms $A$ and $B$ who select "factories." If 170 out of 460 people who used version $A$ chose "factories" and 141 out of 440 people who used version $B$ chose "factories,"

**(a)** Find a 95% confidence interval for $p_A - p_B$.

**(b)** Do the versions seem to be consistent concerning this answer? Why or why not?

## 7.4 SAMPLE SIZE

In statistical consulting, the first question frequently asked is, "How large should the sample size be to estimate a mean?" In order to convince the inquirer that the answer will depend on the variation associated with the random variable under observation, the statistician could correctly respond, "Only one observation is needed, provided that the standard deviation of the distribution is zero." That is, if $\sigma$ equals zero, then the value of that one observation would necessarily equal the unknown mean of the distribution. This, of course, is an extreme case and one that is not met in practice; however, it should help convince people that the smaller the variance, the smaller is the sample size needed to achieve a given degree of accuracy. This assertion will become clearer as we consider several examples. Let us begin with a problem that involves a statistical inference about the unknown mean of a distribution.

**Example 7.4-1** A mathematics department wishes to evaluate a new method of teaching calculus with a computer. At the end of the course, the evaluation will be made on the basis of scores of the participating students on a standard test. There is particular interest in estimating $\mu$, the mean score for students taking the course. Thus, there is a desire to determine the number of students, $n$, who are to be selected at random from a larger group of students to take the course. Since new computing equipment must be purchased, the department cannot afford to let all of the school's students take calculus the new way. In addition, some of the staff question the value of this approach and hence do not want to expose every student to this new procedure. So, let us find the sample size $n$ such that we are fairly confident that $\bar{x} \pm 1$ contains the unknown test mean $\mu$. From past experience, it is believed that the standard deviation associated with this type of test is about 15. (The mean is also known when students take the standard calculus course.) Accordingly, using the fact that the sample mean of the test scores, $\overline{X}$, is approximately $N(\mu, \sigma^2/n)$, we see that the interval given by $\bar{x} \pm 1.96(15/\sqrt{n})$ will serve as an approximate 95% confidence interval for $\mu$. That is, we want

$$1.96\left(\frac{15}{\sqrt{n}}\right) = 1$$

or, equivalently,

$$\sqrt{n} = 29.4, \qquad \text{and thus} \qquad n \approx 864.36,$$

or $n = 865$ because $n$ must be an integer. ■

It is quite likely that, in the preceding example, it had not been anticipated that as many as 865 students would be needed in this study. If that is the case, the statistician must discuss with those involved in the experiment whether or not the accuracy and the confidence level could be relaxed some. For example, rather than

requiring $\bar{x} \pm 1$ to be a 95% confidence interval for $\mu$, possibly $\bar{x} \pm 2$ would be a satisfactory 80% one. If this modification is acceptable, we now have

$$1.282\left(\frac{15}{\sqrt{n}}\right) = 2$$

or, equivalently,

$$\sqrt{n} = 9.615, \qquad \text{so that} \qquad n \approx 92.4.$$

Since $n$ must be an integer, we would probably use 93 in practice. Most likely, the persons involved in the project would find that a more reasonable sample size. Of course, any sample size greater than 93 could be used. Then either the length of the confidence interval could be decreased from $\bar{x} \pm 2$ or the confidence coefficient could be increased from 80%, or a combination of both approaches could be taken. Also, since there might be some question as to whether the standard deviation $\sigma$ actually equals 15, the sample standard deviation $s$ would no doubt be used in the construction of the interval. For instance, suppose that the sample characteristics observed are

$$n = 145, \qquad \bar{x} = 77.2, \qquad s = 13.2;$$

then

$$\bar{x} \pm \frac{1.282s}{\sqrt{n}}, \qquad \text{or} \qquad 77.2 \pm 1.41,$$

provides an approximate 80% confidence interval for $\mu$.

In general, if we want the $100(1-\alpha)\%$ confidence interval for $\mu, \bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$, to be no longer than that given by $\bar{x} \pm \varepsilon$, then the sample size $n$ is the solution of

$$\varepsilon = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \qquad \text{where } \Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}.$$

That is,

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{\varepsilon^2}, \tag{7.4-1}$$

where it is assumed that $\sigma^2$ is known. We sometimes call $\varepsilon = z_{\alpha/2}(\sigma/\sqrt{n})$ the **maximum error of the estimate**. If the experimenter has no idea about the value of $\sigma^2$, it may be necessary to first take a preliminary sample to estimate $\sigma^2$.

The type of statistic we see most often in newspapers and magazines is an estimate of a proportion $p$. We might, for example, want to know the percentage of the labor force that is unemployed or the percentage of voters favoring a certain candidate. Sometimes extremely important decisions are made on the basis of these estimates. If this is the case, we would most certainly desire short confidence intervals for $p$ with large confidence coefficients. We recognize that these conditions will require a large sample size. If, to the contrary, the fraction $p$ being estimated is not too important, an estimate associated with a longer confidence interval with a smaller confidence coefficient is satisfactory, and in that case a smaller sample size can be used.

**Example 7.4-2** Suppose we know that the unemployment rate has been about 8% (0.08). However, we wish to update our estimate in order to make an important decision about the national economic policy. Accordingly, let us say we wish to be 99% confident that

the new estimate of $p$ is within 0.001 of the true $p$. If we assume Bernoulli trials (an assumption that might be questioned), the relative frequency $y/n$, based upon a large sample size $n$, provides the approximate 99% confidence interval:

$$\frac{y}{n} \pm 2.576\sqrt{\frac{(y/n)(1-y/n)}{n}}.$$

Although we do not know $y/n$ exactly before sampling, since $y/n$ will be near 0.08, we do know that

$$2.576\sqrt{\frac{(y/n)(1-y/n)}{n}} \approx 2.576\sqrt{\frac{(0.08)(0.92)}{n}},$$

and we want this number to equal 0.001. That is,

$$2.576\sqrt{\frac{(0.08)(0.92)}{n}} = 0.001$$

or, equivalently,

$$\sqrt{n} = 2576\sqrt{0.0736}, \qquad \text{and then} \qquad n \approx 488{,}394.$$

That is, under our assumptions, such a sample size is needed in order to achieve the reliability and the accuracy desired. Because $n$ is so large, we would probably be willing to increase the error, say, to 0.01, and perhaps reduce the confidence level to 98%. In that case,

$$\sqrt{n} = (2.326/0.01)\sqrt{0.0736} \qquad \text{and} \qquad n \approx 3{,}982,$$

which is a more reasonable sample size. ■

From the preceding example, we hope that the student will recognize how important it is to know the sample size (or the length of the confidence interval and the confidence coefficient) before he or she can place much weight on a statement such as "Fifty-one percent of the voters seem to favor candidate A, 46% favor candidate B, and 3% are undecided." Is this statement based on a sample of 100 or 2000 or 10,000 voters? If we assume Bernoulli trials, the approximate 95% confidence intervals for the fraction of voters favoring candidate A in these cases are, respectively, $[0.41, 0.61]$, $[0.49, 0.53]$, and $[0.50, 0.52]$. Quite obviously, the first interval, with $n = 100$, does not assure candidate A of the support of at least half the voters, whereas the interval with $n = 10{,}000$ is more convincing.

In general, to find the required sample size to estimate $p$, recall that the point estimate of $p$ is $\widehat{p} = y/n$ and an approximate $1-\alpha$ confidence interval for $p$ is

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}\,(1-\widehat{p}\,)}{n}}.$$

Suppose we want an estimate of $p$ that is within $\varepsilon$ of the unknown $p$ with $100(1-\alpha)\%$ confidence, where $\varepsilon = z_{\alpha/2}\sqrt{\widehat{p}\,(1-\widehat{p}\,)/n}$ is the **maximum error of the point estimate** $\widehat{p} = y/n$. Since $\widehat{p}$ is unknown before the experiment is run, we cannot use the value of $\widehat{p}$ in our determination of $n$. However, if it is known that $p$ is about equal to $p^*$, the necessary sample size $n$ is the solution of

$$\varepsilon = \frac{z_{\alpha/2}\sqrt{p^*(1-p^*)}}{\sqrt{n}}.$$

That is,

$$n = \frac{z_{\alpha/2}^2 p^*(1-p^*)}{\varepsilon^2}. \tag{7.4-2}$$

Often, however, we do not have a strong prior idea about $p$, as we did in Example 7.4-2 about the rate of unemployment. It is interesting to observe that no matter what value $p$ takes between 0 and 1, it is always true that $p^*(1-p^*) \le 1/4$. Hence,

$$n = \frac{z_{\alpha/2}^2 p^*(1-p^*)}{\varepsilon^2} \le \frac{z_{\alpha/2}^2}{4\varepsilon^2}.$$

Thus, if we want the $100(1-\alpha)\%$ confidence interval for $p$ to be no longer than $y/n \pm \varepsilon$, a solution for $n$ that provides this protection is

$$n = \frac{z_{\alpha/2}^2}{4\varepsilon^2}. \tag{7.4-3}$$

REMARK Up to this point in the text, we have used the "hat" ($\widehat{\ }$) notation to indicate an estimator, as in $\widehat{p} = Y/n$ and $\widehat{\mu} = \overline{X}$. Note, however, that in the previous discussion we used $\widehat{p} = y/n$, an estimate of $p$. Occasionally, statisticians find it convenient to use the "hat" notation for an estimate as well as an estimator. It is usually clear from the context which is being used. ■

**Example 7.4-3**

A possible gubernatorial candidate wants to assess initial support among the voters before making an announcement about her candidacy. If the fraction $p$ of voters who are favorable, without any advance publicity, is around 0.15, the candidate will enter the race. From a poll of $n$ voters selected at random, the candidate would like the estimate $y/n$ to be within 0.03 of $p$. That is, the decision will be based on a 95% confidence interval of the form $y/n \pm 0.03$. Since the candidate has no idea about the magnitude of $p$, a consulting statistician formulates the equation

$$n = \frac{(1.96)^2}{4(0.03)^2} = 1067.11.$$

Thus, the sample size should be around 1068 to achieve the desired reliability and accuracy. Suppose that 1068 voters around the state were selected at random and interviewed and $y = 214$ express support for the candidate. Then $\widehat{p} = 214/1068 = 0.20$ is a point estimate of $p$, and an approximate 95% confidence interval for $p$ is

$$0.20 \pm 1.96\sqrt{(0.20)(0.80)/n}, \qquad \text{or} \qquad 0.20 \pm 0.024.$$

That is, we are 95% confident that $p$ belongs to the interval $[0.176, 0.224]$. On the basis of this sample, the candidate decided to run for office. Note that, for a confidence coefficient of 95%, we found a sample size so that the maximum error of the estimate would be 0.03. From the data that were collected, the maximum error of the estimate is only 0.024. We ended up with a smaller error because we found the sample size assuming that $p = 0.50$, while, in fact, $p$ is closer to 0.20. ■

Suppose that you want to estimate the proportion $p$ of a student body that favors a new policy. How large should the sample be? If $p$ is close to 1/2 and you want to be 95% confident that the maximum error of the estimate is $\varepsilon = 0.02$, then

$$n = \frac{(1.96)^2}{4(0.02)^2} = 2401.$$

Such a sample size makes sense at a large university. However, if you are a student at a small college, the entire enrollment could be less than 2401. Thus, we now give a procedure that can be used to determine the sample size when the population is not so large relative to the desired sample size.

Let $N$ equal the size of a population, and assume that $N_1$ individuals in the population have a certain characteristic $C$ (e.g., favor a new policy). Let $p = N_1/N$, the proportion with this characteristic. Then $1-p = 1-N_1/N$. If we take a sample of size $n$ without replacement, then $X$, the number of observations with the characteristic $C$, has a hypergeometric distribution. The mean and variance of $X$ are, respectively,

$$\mu = n\left(\frac{N_1}{N}\right) = np$$

and

$$\sigma^2 = n\left(\frac{N_1}{N}\right)\left(1 - \frac{N_1}{N}\right)\left(\frac{N-n}{N-1}\right) = np(1-p)\left(\frac{N-n}{N-1}\right).$$

The mean and variance of $X/n$ are, respectively,

$$E\left(\frac{X}{n}\right) = \frac{\mu}{n} = p$$

and

$$\text{Var}\left(\frac{X}{n}\right) = \frac{\sigma^2}{n^2} = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right).$$

To find an approximate confidence interval for $p$, we can use the normal approximation:

$$P\left[-z_{\alpha/2} \le \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}} \le z_{\alpha/2}\right] \approx 1 - \alpha.$$

Thus,

$$1 - \alpha \approx P\left[\frac{X}{n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)} \le p \le \frac{X}{n} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}\right].$$

Replacing $p$ under the radical with $\hat{p} = x/n$, we find that an approximate $1 - \alpha$ confidence interval for $p$ is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p}\,)}{n}\left(\frac{N-n}{N-1}\right)}.$$

This is similar to the confidence interval for $p$ when the distribution of $X$ is $b(n,p)$. If $N$ is large relative to $n$, then

$$\frac{N-n}{N-1} = \frac{1-n/N}{1-1/N} \approx 1,$$

so in this case the two intervals are essentially equal.

Suppose now that we are interested in determining the sample size $n$ that is required to have $1-\alpha$ confidence that the maximum error of the estimate of $p$ is $\varepsilon$. We let

$$\varepsilon = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}$$

and solve for $n$. After some simplification, we obtain

$$n = \frac{Nz_{\alpha/2}^2 p(1-p)}{(N-1)\varepsilon^2 + z_{\alpha/2}^2 p(1-p)}$$

$$= \frac{z_{\alpha/2}^2 p(1-p)/\varepsilon^2}{\dfrac{N-1}{N} + \dfrac{z_{\alpha/2}^2 p(1-p)/\varepsilon^2}{N}}.$$

If we let

$$m = \frac{z_{\alpha/2}^2 p^*(1-p^*)}{\varepsilon^2},$$

which is the $n$ value given by Equation 7.4-2, then we choose

$$n = \frac{m}{1+\dfrac{m-1}{N}}$$

for our sample size $n$.

If we know nothing about $p$, we set $p^* = 1/2$ to determine $m$. For example, if the size of the student body is $N = 4000$ and $1-\alpha = 0.95$, $\varepsilon = 0.02$, and we let $p^* = 1/2$, then $m = 2401$ and

$$n = \frac{2401}{1+2400/4000} = 1501,$$

rounded up to the nearest integer. Thus, we would sample approximately 37.5% of the student body.

**Example 7.4-4** Suppose that a college of $N = 3000$ students is interested in assessing student support for a new form for teacher evaluation. To estimate the proportion $p$ in favor of the new form, how large a sample is required so that the maximum error of the estimate of $p$ is $\varepsilon = 0.03$ with 95% confidence? If we assume that $p$ is completely unknown, we use $p^* = 1/2$ to obtain

$$m = \frac{(1.96)^2}{4(0.03)^2} = 1068,$$

rounded up to the nearest integer. Thus, the desired sample size is

$$n = \frac{1068}{1 + 1067/3000} = 788,$$

rounded up to the nearest integer. ■

## Exercises

**7.4-1.** Let $X$ equal the tarsus length for a male grackle. Assume that the distribution of $X$ is $N(\mu, 4.84)$. Find the sample size $n$ that is needed so that we are 95% confident that the maximum error of the estimate of $\mu$ is 0.4.

**7.4-2.** Let $X$ equal the excess weight of soap in a "1000-gram" bottle. Assume that the distribution of $X$ is $N(\mu, 169)$. What sample size is required so that we have 95% confidence that the maximum error of the estimate of $\mu$ is 1.5?

**7.4-3.** A company packages powdered soap in "6-pound" boxes. The sample mean and standard deviation of the soap in these boxes are currently 6.09 pounds and 0.02 pound, respectively. If the mean fill can be lowered by 0.01 pound, \$14,000 would be saved per year. Adjustments were made in the filling equipment, but it can be assumed that the standard deviation remains unchanged.

**(a)** How large a sample is needed so that the maximum error of the estimate of the new $\mu$ is $\varepsilon = 0.001$ with 90% confidence?

**(b)** A random sample of size $n = 1219$ yielded $\bar{x} = 6.048$ and $s = 0.022$. Calculate a 90% confidence interval for $\mu$.

**(c)** Estimate the savings per year with these new adjustments.

**(d)** Estimate the proportion of boxes that will now weigh less than 6 pounds.

**7.4-4.** Measurements of the length in centimeters of $n = 29$ fish yielded an average length of $\bar{x} = 16.82$ and $s^2 = 34.9$. Determine the size of a new sample so that $\bar{x} \pm 0.5$ is an approximate 95% confidence interval for $\mu$.

**7.4-5.** A quality engineer wanted to be 98% confident that the maximum error of the estimate of the mean strength, $\mu$, of the left hinge on a vanity cover molded by a machine is 0.25. A preliminary sample of size $n = 32$ parts yielded a sample mean of $\bar{x} = 35.68$ and a standard deviation of $s = 1.723$.

**(a)** How large a sample is required?

**(b)** Does this seem to be a reasonable sample size? (Note that destructive testing is needed to obtain the data.)

**7.4-6.** A manufacturer sells a light bulb that has a mean life of 1450 hours with a standard deviation of 33.7 hours. A new manufacturing process is being tested, and there is interest in knowing the mean life $\mu$ of the new bulbs. How large a sample is required so that $\bar{x} \pm 5$ is a 95% confidence interval for $\mu$? You may assume that the change in the standard deviation is minimal.

**7.4-7.** For a public opinion poll for a close presidential election, let $p$ denote the proportion of voters who favor candidate $A$. How large a sample should be taken if we want the maximum error of the estimate of $p$ to be equal to

**(a)** 0.03 with 95% confidence?

**(b)** 0.02 with 95% confidence?

**(c)** 0.03 with 90% confidence?

**7.4-8.** Some college professors and students examined 137 Canadian geese for patent schistosome in the year they hatched. Of these 137 birds, 54 were infected. The professors and students were interested in estimating $p$, the proportion of infected birds of this type. For future studies, determine the sample size $n$ so that the estimate of $p$ is within $\varepsilon = 0.04$ of the unknown $p$ with 90% confidence.

**7.4-9.** A die has been loaded to change the probability of rolling a 6. In order to estimate $p$, the new probability of rolling a 6, how many times must the die be rolled so that we are 99% confident that the maximum error of the estimate of $p$ is $\varepsilon = 0.02$?

**7.4-10.** A seed distributor claims that 80% of its beet seeds will germinate. How many seeds must be tested for germination in order to estimate $p$, the true proportion that will germinate, so that the maximum error of the estimate is $\varepsilon = 0.03$ with 90% confidence?

**7.4-11.** Some dentists were interested in studying the fusion of embryonic rat palates by a standard transplantation technique. When no treatment is used, the probability of fusion equals approximately 0.89. The dentists would like to estimate $p$, the probability of fusion, when vitamin A is lacking.

**(a)** How large a sample $n$ of rat embryos is needed for $y/n \pm 0.10$ to be a 95% confidence interval for $p$?

**(b)** If $y = 44$ out of $n = 60$ palates showed fusion, give a 95% confidence interval for $p$.

**7.4-12.** Let $p$ equal the proportion of college students who favor a new policy for alcohol consumption on campus. How large a sample is required to estimate $p$ so that the maximum error of the estimate of $p$ is 0.04 with 95% confidence when the size of the student body is

**(a)** $N = 1500$?

**(b)** $N = 15{,}000$?

**(c)** $N = 25{,}000$?

**7.4-13.** Out of 1000 welds that have been made on a tower, it is suspected that 15% are defective. To estimate $p$, the proportion of defective welds, how many welds must be inspected to have approximately 95% confidence that the maximum error of the estimate of $p$ is 0.04?

**7.4-14.** If $Y_1/n$ and $Y_2/n$ are the respective independent relative frequencies of success associated with the two binomial distributions $b(n,p_1)$ and $b(n,p_2)$, compute $n$ such that the approximate probability that the random interval $(Y_1/n - Y_2/n) \pm 0.05$ covers $p_1 - p_2$ is at least 0.80. HINT: Take $p_1^* = p_2^* = 1/2$ to provide an upper bound for $n$.

**7.4-15.** If $\overline{X}$ and $\overline{Y}$ are the respective means of two independent random samples of the same size $n$, find $n$ if we want $\bar{x} - \bar{y} \pm 4$ to be a 90% confidence interval for $\mu_X - \mu_Y$. Assume that the standard deviations are known to be $\sigma_X = 15$ and $\sigma_Y = 25$.

## 7.5 DISTRIBUTION-FREE CONFIDENCE INTERVALS FOR PERCENTILES

In Section 6.3, we defined sample percentiles in terms of order statistics and noted that the sample percentiles can be used to estimate corresponding distribution percentiles. In this section, we use order statistics to construct confidence intervals for unknown distribution percentiles. Since little is assumed about the underlying distribution (except that it is of the continuous type) in the construction of these confidence intervals, they are often called **distribution-free confidence intervals**.

If $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ are the order statistics of a random sample of size $n = 5$ from a continuous-type distribution, then the sample median $Y_3$ could be thought of as an estimator of the distribution median $\pi_{0.5}$. We shall let $m = \pi_{0.5}$. We could simply use the sample median $Y_3$ as an estimator of the distribution median $m$. However, we are certain that all of us recognize that, with only a sample of size 5, we would be quite lucky if the observed $Y_3 = y_3$ were very close to $m$. Thus, we now describe how a confidence interval can be constructed for $m$.

Instead of simply using $Y_3$ as an estimator of $m$, let us also compute the probability that the random interval $(Y_1, Y_5)$ includes $m$. That is, let us determine $P(Y_1 < m < Y_5)$. Doing this is easy if we say that we have success if an individual observation—say, $X$—is less than $m$; then the probability of success on one of the independent trials is $P(X < m) = 0.5$. In order for the first order statistic $Y_1$ to be less than $m$ and the last order statistic $Y_5$ to be greater than $m$, we must have at least one success, but not five successes. That is,

$$P(Y_1 < m < Y_5) = \sum_{k=1}^{4} \binom{5}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{5-k}$$

$$= 1 - \left(\frac{1}{2}\right)^5 - \left(\frac{1}{2}\right)^5 = \frac{15}{16}.$$

So the probability that the random interval $(Y_1, Y_5)$ includes $m$ is $15/16 \approx 0.94$. Suppose now that this random sample is actually taken and the order statistics are observed to equal $y_1 < y_2 < y_3 < y_4 < y_5$, respectively. Then $(y_1, y_5)$ is a 94% confidence interval for $m$.

It is interesting to note what happens as the sample size increases. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size $n$ from a distribution of the continuous type. Then $P(Y_1 < m < Y_n)$ is the probability that there is at least

one "success" but not $n$ successes, where the probability of success on each trial is $P(X < m) = 0.5$. Consequently,

$$P(Y_1 < m < Y_n) = \sum_{k=1}^{n-1} \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k}$$
$$= 1 - \left(\frac{1}{2}\right)^n - \left(\frac{1}{2}\right)^n = 1 - \left(\frac{1}{2}\right)^{n-1}.$$

This probability increases as $n$ increases, so that the corresponding confidence interval $(y_1, y_n)$ would have the very large confidence coefficient $1 - (1/2)^{n-1}$. Unfortunately, the interval $(y_1, y_n)$ tends to get wider as $n$ increases; thus, we are not "pinning down" $m$ very well. However, if we used the interval $(y_2, y_{n-1})$ or $(y_3, y_{n-2})$, we would obtain shorter intervals, but also smaller confidence coefficients. Let us investigate this possibility further.

With the order statistics $Y_1 < Y_2 < \cdots < Y_n$ associated with a random sample of size $n$ from a continuous-type distribution, consider $P(Y_i < m < Y_j)$, where $i < j$. For example, we might want

$$P(Y_2 < m < Y_{n-1}) \qquad \text{or} \qquad P(Y_3 < m < Y_{n-2}).$$

On each of the $n$ independent trials, we say that we have success if that $X$ is less than $m$; thus, the probability of success on each trial is $P(X < m) = 0.5$. Consequently, to have the $i$th order statistic $Y_i$ less than $m$ and the $j$th order statistic greater than $m$, we must have at least $i$ successes but fewer than $j$ successes (or else $Y_j < m$). That is,

$$P(Y_i < m < Y_j) = \sum_{k=i}^{j-1} \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} = 1 - \alpha.$$

For particular values of $n$, $i$, and $j$, this probability—say, $1 - \alpha$—which is the sum of probabilities from a binomial distribution, can be calculated directly or approximated by an area under the normal pdf, provided that $n$ is large enough. The observed interval $(y_i, y_j)$ could then serve as a $100(1 - \alpha)\%$ confidence interval for the unknown distribution median.

**Example 7.5-1**

The lengths in centimeters of $n = 9$ fish of a particular species captured off the New England coast were 32.5, 27.6, 29.3, 30.1, 15.5, 21.7, 22.8, 21.2, and 19.0. Thus, the observed order statistics are

$$15.5 < 19.0 < 21.2 < 21.7 < 22.8 < 27.6 < 29.3 < 30.1 < 32.5.$$

Before the sample is drawn, we know that

$$P(Y_2 < m < Y_8) = \sum_{k=2}^{7} \binom{9}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{9-k}$$
$$= 0.9805 - 0.0195 = 0.9610,$$

from Table II in Appendix B. Thus, the confidence interval $(y_2 = 19.0, y_8 = 30.1)$ for $m$, the median of the lengths of all fish of this species, has a 96.1% confidence coefficient. ■

So that the student need not compute many of these probabilities, Table 7.5-1 lists the necessary information for constructing confidence intervals of the form $(y_i, y_{n+1-i})$ for the unknown $m$ for sample sizes $n = 5, 6, \ldots, 20$. The subscript $i$ is selected so that the confidence coefficient $P(Y_i < m < Y_{n+1-i})$ is greater than 90% and as close to 95% as possible.

For sample sizes larger than 20, we approximate those binomial probabilities with areas under the normal curve. To illustrate how good these approximations are, we compute the probability corresponding to $n = 16$ in Table 7.5-1. Here, using Table II, we have

$$\begin{aligned} 1-\alpha &= P(Y_5 < m < Y_{12}) = \sum_{k=5}^{11} \binom{16}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{16-k} \\ &= P(W = 5, 6, \ldots, 11) \\ &= 0.9616 - 0.0384 = 0.9232, \end{aligned}$$

where $W$ is $b(16, 1/2)$. The normal approximation gives

$$1-\alpha = P(4.5 < W < 11.5) = P\left(\frac{4.5-8}{2} < \frac{W-8}{2} < \frac{11.5-8}{2}\right),$$

because $W$ has mean $np = 8$ and variance $np(1-p) = 4$. The standardized variable $Z = (W-8)/2$ has an approximate normal distribution. Thus,

$$\begin{aligned} 1-\alpha &\approx \Phi\left(\frac{3.5}{2}\right) - \Phi\left(\frac{-3.5}{2}\right) = \Phi(1.75) - \Phi(-1.75) \\ &= 0.9599 - 0.0401 = 0.9198. \end{aligned}$$

This value compares very favorably with the probability 0.9232 recorded in Table 7.5-1. (Note that Minitab or some other computer program can also be used.)

**Table 7.5-1** Information for confidence intervals for $m$

| $n$ | $(i, n+1-i)$ | $P(Y_i < m < Y_{n+1-i})$ | $n$ | $(i, n+1-i)$ | $P(Y_i < m < Y_{n+1-i})$ |
|---|---|---|---|---|---|
| 5 | (1, 5) | 0.9376 | 13 | (3, 11) | 0.9776 |
| 6 | (1, 6) | 0.9688 | 14 | (4, 11) | 0.9426 |
| 7 | (1, 7) | 0.9844 | 15 | (4, 12) | 0.9648 |
| 8 | (2, 7) | 0.9296 | 16 | (5, 12) | 0.9232 |
| 9 | (2, 8) | 0.9610 | 17 | (5, 13) | 0.9510 |
| 10 | (2, 9) | 0.9786 | 18 | (5, 14) | 0.9692 |
| 11 | (3, 9) | 0.9346 | 19 | (6, 14) | 0.9364 |
| 12 | (3, 10) | 0.9614 | 20 | (6, 15) | 0.9586 |

The argument used to find a confidence interval for the median $m$ of a distribution of the continuous type can be applied to any percentile $\pi_p$. In this case, we say that we have success on a single trial if that $X$ is less than $\pi_p$. Thus, the probability of success on each of the independent trials is $P(X < \pi_p) = p$. Accordingly, with $i < j$, $1-\alpha = P(Y_i < \pi_p < Y_j)$ is the probability that we have at least $i$ successes but fewer than $j$ successes. Hence,

$$1 - \alpha = P(Y_i < \pi_p < Y_j) = \sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k}.$$

Once the sample is observed and the order statistics determined, the known interval $(y_i, y_j)$ could serve as a $100(1-\alpha)\%$ confidence interval for the unknown distribution percentile $\pi_p$.

**Example 7.5-2**

Let the following numbers represent the order statistics of the $n = 27$ observations obtained in a random sample from a certain population of incomes (measured in hundreds of dollars):

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 261 | 269 | 271 | 274 | 279 | 280 | 283 | 284 | 286 |
| 287 | 292 | 293 | 296 | 300 | 304 | 305 | 313 | 321 |
| 322 | 329 | 341 | 343 | 356 | 364 | 391 | 417 | 476 |

Say we are interested in estimating the 25th percentile, $\pi_{0.25}$, of the population. Since $(n+1)p = 28(1/4) = 7$, the seventh order statistic, namely, $y_7 = 283$, would be a point estimate of $\pi_{0.25}$. To find a confidence interval for $\pi_{0.25}$, let us move down and up a few order statistics from $y_7$—say, to $y_4$ and $y_{10}$. What is the confidence coefficient associated with the interval $(y_4, y_{10})$? Before the sample was drawn, we had

$$1 - \alpha = P(Y_4 < \pi_{0.25} < Y_{10}) = \sum_{k=4}^{9} \binom{27}{k} (0.25)^k (0.75)^{27-k} = 0.8201.$$

For the normal approximation, we use $W$, which is $b(27, 1/4)$ with mean $27/4 = 6.75$ and variance $81/16$. Hence,

$$\begin{aligned} 1 - \alpha &= P(4 \le W \le 9) = P(3.5 < W < 9.5) \\ &\approx \Phi\left(\frac{9.5 - 6.75}{9/4}\right) - \Phi\left(\frac{3.5 - 6.75}{9/4}\right) \\ &= \Phi\left(\frac{11}{9}\right) - \Phi\left(-\frac{13}{9}\right) = 0.8149. \end{aligned}$$

Thus, $(y_4 = 274, y_{10} = 287)$ is an 82.01% (or approximate 81.49%) confidence interval for $\pi_{0.25}$. Note that we could choose other intervals, such as $(y_3 = 271, y_{11} = 292)$, and these would have different confidence coefficients. The persons involved in the study must select the desired confidence coefficient, and then the appropriate order statistics are taken, usually quite symmetrically about the $(n+1)p$th order statistic. ■

When the number of observations is large, it is important to be able to determine the order statistics rather easily. As illustrated in the next example, a stem-and-leaf diagram, as introduced in Section 6.2, can be helpful in determining the needed order statistics.

**Example 7.5-3**

The measurements of butterfat produced by $n = 90$ cows during a 305-day milk production period following their first calf are summarized in Table 7.5-2, in which each leaf consists of two digits. From this display, it is quite easy to see that $y_8 = 392$.

**Table 7.5-2** Ordered stem-and-leaf diagram of butterfat production

| Stems | Leaves | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **2s** | 74 | | | | | | | | | |
| **2•** | | | | | | | | | | |
| **3∗** | | | | | | | | | | |
| **3t** | 27 | 39 | | | | | | | | |
| **3f** | 45 | 50 | | | | | | | | |
| **3s** | | | | | | | | | | |
| **3•** | 80 | 88 | 92 | 94 | 95 | | | | | |
| **4∗** | 17 | 18 | | | | | | | | |
| **4t** | 21 | 22 | 27 | 34 | 37 | 39 | | | | |
| **4f** | 44 | 52 | 53 | 53 | 57 | 58 | | | | |
| **4s** | 60 | 64 | 66 | 70 | 70 | 72 | 75 | 78 | | |
| **4•** | 81 | 86 | 89 | 91 | 92 | 94 | 96 | 97 | 99 | |
| **5∗** | 00 | 00 | 01 | 02 | 05 | 09 | 10 | 13 | 13 | 16 |
| **5t** | 24 | 26 | 31 | 32 | 32 | 37 | 37 | 39 | | |
| **5f** | 40 | 41 | 44 | 55 | | | | | | |
| **5s** | 61 | 70 | 73 | 74 | | | | | | |
| **5•** | 83 | 83 | 86 | 93 | 99 | | | | | |
| **6∗** | 07 | 08 | 11 | 12 | 13 | 17 | 18 | 19 | | |
| **6t** | 27 | 28 | 35 | 37 | | | | | | |
| **6f** | 43 | 43 | 45 | | | | | | | |
| **6s** | 72 | | | | | | | | | |
| **6•** | 91 | 96 | | | | | | | | |

It takes a little more work to show that $y_{38} = 494$ and $y_{53} = 526$ creates an interval $(494, 526)$ which serves as a confidence interval for the unknown median $m$ of all butterfat production for the given breed of cows. Its confidence coefficient is

$$\begin{aligned} P(Y_{38} < m < Y_{53}) &= \sum_{k=38}^{52} \binom{90}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{90-k} \\ &\approx \Phi\left(\frac{52.5 - 45}{\sqrt{22.5}}\right) - \Phi\left(\frac{37.5 - 45}{\sqrt{22.5}}\right) \\ &= \Phi(1.58) - \Phi(-1.58) \;=\; 0.8858. \end{aligned}$$

Similarly, $(y_{17} = 437, y_{29} = 470)$ is a confidence interval for the first quartile, $\pi_{0.25}$, with confidence coefficient

$$\begin{aligned} P(Y_{17} < \pi_{0.25} < Y_{29}) &\approx \Phi\left(\frac{28.5 - 22.5}{\sqrt{16.875}}\right) - \Phi\left(\frac{16.5 - 22.5}{\sqrt{16.875}}\right) \\ &= \Phi(1.46) - \Phi(-1.46) = 0.8558. \end{aligned}$$

Using the binomial distribution, the confidence coefficients are 0.8867 and 0.8569, respectively. ■

It is interesting to compare the length of a confidence interval for the mean $\mu$ obtained with $\bar{x} \pm t_{\alpha/2}(n-1)(s/\sqrt{n})$ against the length of a $100(1-\alpha)\%$ confidence interval for the median $m$ obtained with the distribution-free techniques of this section. Usually, if the sample arises from a distribution that does not deviate too much from the normal, the confidence interval based upon $\bar{x}$ is much shorter. After all, we assume much more when we create that confidence interval. With the distribution-free method, all we assume is that the distribution is of the continuous type. So if the distribution is highly skewed or heavy-tailed so that outliers could exist, a distribution-free technique is safer and much more robust. Moreover, the distribution-free technique provides a way to get confidence intervals for various percentiles, and investigators are often interested in such intervals.

## Exercises

**7.5-1.** Let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5 < Y_6$ be the order statistics of a random sample of size $n = 6$ from a distribution of the continuous type having $(100p)$th percentile $\pi_p$. Compute

**(a)** $P(Y_2 < \pi_{0.5} < Y_5)$.

**(b)** $P(Y_1 < \pi_{0.25} < Y_4)$.

**(c)** $P(Y_4 < \pi_{0.9} < Y_6)$.

**7.5-2.** For $n = 12$ year-2007 model sedans whose horsepower is between 290 and 390, the following measurements give the time in seconds for the car to go from 0 to 60 mph:

6.0 6.3 5.0 6.0 5.7 5.9 6.8 5.5 5.4 4.8 5.4 5.8

**(a)** Find a 96.14% confidence interval for the median, $m$.

**(b)** The interval $(y_1, y_7)$ could serve as a confidence interval for $\pi_{0.3}$. Find it and give its confidence coefficient.

**7.5-3.** A sample of $n = 9$ electrochromic mirrors was used to measure the following low-end reflectivity percentages:

7.12 7.22 6.78 6.31 5.99 6.58 7.80 7.40 7.05

**(a)** Find the endpoints for an approximate 95% confidence interval for the median, $m$.

**(b)** The interval $(y_3, y_7)$ could serve as a confidence interval for $m$. Find it and give its confidence coefficient.

**7.5-4.** Let $m$ denote the median weight of "80-pound" bags of water softener pellets. Use the following random sample of $n = 14$ weights to find an approximate 95% confidence interval for $m$:

| | | | | | | |
|---|---|---|---|---|---|---|
| 80.51 | 80.28 | 80.40 | 80.35 | 80.38 | 80.28 | 80.27 |
| 80.16 | 80.59 | 80.56 | 80.32 | 80.27 | 80.53 | 80.32 |

**(a)** Find a 94.26% confidence interval for $m$.

**(b)** The interval $(y_6, y_{12})$ could serve as a confidence interval for $\pi_{0.6}$. What is its confidence coefficient?

**7.5-5.** A biologist who studies spiders selected a random sample of 20 male green lynx spiders (a spider that does not weave a web, but chases and leaps on its prey) and measured the lengths (in millimeters) of one of the front legs of the 20 spiders. Use the following measurements to construct a confidence interval for $m$ that has a confidence coefficient about equal to 0.95:

| | | | | |
|---|---|---|---|---|
| 15.10 | 13.55 | 15.75 | 20.00 | 15.45 |
| 13.60 | 16.45 | 14.05 | 16.95 | 19.05 |
| 16.40 | 17.05 | 15.25 | 16.65 | 16.25 |
| 17.75 | 15.40 | 16.80 | 17.55 | 19.05 |

**7.5-6.** A company manufactures mints that have a label weight of 20.4 grams. The company regularly samples from the production line and weighs the selected mints. During two mornings of production it sampled 81 mints, obtaining the following weights:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 21.8 | 21.7 | 21.7 | 21.6 | 21.3 | 21.6 | 21.5 | 21.3 | 21.2 |
| 21.0 | 21.6 | 21.6 | 21.6 | 21.5 | 21.4 | 21.8 | 21.7 | 21.6 |
| 21.6 | 21.3 | 21.9 | 21.9 | 21.6 | 21.0 | 20.7 | 21.8 | 21.7 |
| 21.7 | 21.4 | 20.9 | 22.0 | 21.3 | 21.2 | 21.0 | 21.0 | 21.9 |
| 21.7 | 21.5 | 21.5 | 21.1 | 21.3 | 21.3 | 21.2 | 21.0 | 20.8 |
| 21.6 | 21.6 | 21.5 | 21.5 | 21.2 | 21.5 | 21.4 | 21.4 | 21.3 |
| 21.2 | 21.8 | 21.7 | 21.7 | 21.6 | 20.5 | 21.8 | 21.7 | 21.5 |
| 21.4 | 21.4 | 21.9 | 21.8 | 21.7 | 21.4 | 21.3 | 20.9 | 21.9 |
| 20.7 | 21.1 | 20.8 | 20.6 | 20.6 | 22.0 | 22.0 | 21.7 | 21.6 |

**(a)** Construct an ordered stem-and-leaf display using stems of $20f$, $20s$, $20\bullet$, $21*, \ldots, 22*$.

**(b)** Find **(i)** the three quartiles, **(ii)** the 60th percentile, and **(iii)** the 15th percentile.

**(c)** Find approximate 95% confidence intervals for **(i)** $\pi_{0.25}$, **(ii)** $m = \pi_{0.5}$, and **(iii)** $\pi_{0.75}$.

**7.5-7.** Here are the weights (in grams) of 25 indicator housings used on gauges (see Exercise 6.2-8):

| | | | | |
|---|---|---|---|---|
| 102.0 | 106.3 | 106.6 | 108.8 | 107.7 |
| 106.1 | 105.9 | 106.7 | 106.8 | 110.2 |
| 101.7 | 106.6 | 106.3 | 110.2 | 109.9 |
| 102.0 | 105.8 | 109.1 | 106.7 | 107.3 |
| 102.0 | 106.8 | 110.0 | 107.9 | 109.3 |

**(a)** List the observations in order of magnitude.

**(b)** Give point estimates of $\pi_{0.25}$, $m$, and $\pi_{0.75}$.

**(c)** Find the following confidence intervals and, from Table II in Appendix B, state the associated confidence coefficient:

**(i)** $(y_3, y_{10})$, a confidence interval for $\pi_{0.25}$.

**(ii)** $(y_9, y_{17})$, a confidence interval for the median $m$.

**(iii)** $(y_{16}, y_{23})$, a confidence interval for $\pi_{0.75}$.

**(d)** Use $\bar{x} \pm t_{\alpha/2}(24)(s/\sqrt{25})$ to find a confidence interval for $\mu$, whose confidence coefficient corresponds to that of (c), part (ii). Compare these two confidence intervals of the middles.

**7.5-8.** The biologist of Exercise 7.5-5 also selected a random sample of 20 female green lynx spiders and measured the length (again in millimeters) of one of their front legs. Use the following data to construct a confidence interval for $m$ that has a confidence coefficient about equal to 0.95:

| | | | | |
|---|---|---|---|---|
| 15.85 | 18.00 | 11.45 | 15.60 | 16.10 |
| 18.80 | 12.85 | 15.15 | 13.30 | 16.65 |
| 16.25 | 16.15 | 15.25 | 12.10 | 16.20 |
| 14.80 | 14.60 | 17.05 | 14.15 | 15.85 |

**7.5-9.** Let $X$ equal the amount of fluoride in a certain brand of toothpaste. The specifications are 0.85–1.10 mg/g. Table 6.1-3 lists 100 such measurements.

**(a)** Give a point estimate of the median $m = \pi_{0.50}$.

**(b)** Find an approximate 95% confidence interval for the median $m$. If possible, use a computer to find the exact confidence level.

**(c)** Give a point estimate for the first quartile.

**(d)** Find an approximate 95% confidence interval for the first quartile and, if possible, give the exact confidence coefficient.

**(e)** Give a point estimate for the third quartile.

**(f)** Find an approximate 95% confidence interval for the third quartile and, if possible, give the exact confidence coefficient.

**7.5-10.** When placed in solutions of varying ionic strength, paramecia grow blisters in order to counteract the flow of water. The following 60 measurements in microns are blister lengths:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 7.42 | 5.73 | 3.80 | 5.20 | 11.66 | 8.51 | 6.31 | 8.49 |
| 10.31 | 6.92 | 7.36 | 5.92 | 6.74 | 8.93 | 9.61 | 11.38 |
| 12.78 | 11.43 | 6.57 | 13.50 | 10.58 | 8.03 | 10.07 | 8.71 |
| 10.09 | 11.16 | 7.22 | 10.10 | 6.32 | 10.30 | 10.75 | 11.51 |
| 11.55 | 11.41 | 9.40 | 4.74 | 6.52 | 12.10 | 6.01 | 5.73 |
| 7.57 | 7.80 | 6.84 | 6.95 | 8.93 | 8.92 | 5.51 | 6.71 |
| 10.40 | 13.44 | 9.33 | 8.57 | 7.08 | 8.11 | 13.34 | 6.58 |
| 8.82 | 7.70 | 12.22 | 7.46 | | | | |

**(a)** Construct an ordered stem-and-leaf diagram.

**(b)** Give a point estimate of the median $m = \pi_{0.50}$.

**(c)** Find an approximate 95% confidence interval for $m$.

**(d)** Give a point estimate for the 40th percentile, $\pi_{0.40}$.

**(e)** Find an approximate 90% confidence interval for $\pi_{0.40}$.

**7.5-11.** Using the weights of Verica's 39 gold coins given in Example 6.2-4, find approximate 95% confidence intervals for $\pi_{0.25}$, $\pi_{0.5}$, and $\pi_{0.75}$. Give the exact confidence coefficients for the intervals.

**7.5-12.** Let $Y_1 < Y_2 < \cdots < Y_8$ be the order statistics of eight independent observations from a continuous-type distribution with 70th percentile $\pi_{0.7} = 27.3$.

**(a)** Determine $P(Y_7 < 27.3)$.

**(b)** Find $P(Y_5 < 27.3 < Y_8)$.

## 7.6* MORE REGRESSION

In this section, we develop confidence intervals for important quantities in the linear regression model using the notation and assumptions of Section 6.5. It can be shown (Exercise 7.6-13) that

$$\begin{aligned}\sum_{i=1}^{n}[Y_i-\alpha-\beta(x_i-\bar{x})]^2 &= \sum_{i=1}^{n}\{(\widehat{\alpha}-\alpha)+(\widehat{\beta}-\beta)(x_i-\bar{x})\\ &\quad +[Y_i-\widehat{\alpha}-\widehat{\beta}(x_i-\bar{x})]\}^2\\ &= n(\widehat{\alpha}-\alpha)^2+(\widehat{\beta}-\beta)^2\sum_{i=1}^{n}(x_i-\bar{x})^2\\ &\quad +\sum_{i=1}^{n}[Y_i-\widehat{\alpha}-\widehat{\beta}(x_i-\bar{x})]^2. \qquad (7.6\text{-}1)\end{aligned}$$

From the fact that $Y_i$, $\widehat{\alpha}$, and $\widehat{\beta}$ have normal distributions, it follows that each of

$$\frac{[Y_i-\alpha-\beta(x_i-\bar{x})]^2}{\sigma^2}, \qquad \frac{(\widehat{\alpha}-\alpha)^2}{\left[\frac{\sigma^2}{n}\right]}, \qquad \text{and} \qquad \frac{(\widehat{\beta}-\beta)^2}{\left[\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}$$

has a chi-square distribution with one degree of freedom. Since $Y_1, Y_2, \ldots, Y_n$ are mutually independent,

$$\frac{\sum_{i=1}^{n}[Y_i-\alpha-\beta(x_i-\bar{x})]^2}{\sigma^2}$$

is $\chi^2(n)$. That is, the left-hand member of Equation 7.6-1 divided by $\sigma^2$ is $\chi^2(n)$ and is equal to the sum of two $\chi^2(1)$ variables and

$$\frac{\sum_{i=1}^{n}[Y_i-\widehat{\alpha}-\widehat{\beta}(x_i-\bar{x})]^2}{\sigma^2}=\frac{n\widehat{\sigma^2}}{\sigma^2}\geq 0.$$

Thus, we might guess that $n\widehat{\sigma^2}/\sigma^2$ is $\chi^2(n–2)$. This is true, and moreover, $\widehat{\alpha}$, $\widehat{\beta}$, and $\widehat{\sigma^2}$ are mutually independent. [For a proof, see Hogg, McKean, and Craig, *Introduction to Mathematical Statistics*, 7th ed. (Upper Saddle River, NJ: Prentice Hall, 2013).]

Suppose now that we are interested in forming a confidence interval for $\beta$, the slope of the line. We can use the fact that

$$T_1 = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\left(\frac{\widehat{\beta} - \beta}{\sigma}\right)}{\sqrt{\frac{n\widehat{\sigma^2}}{\sigma^2(n-2)}}} = \frac{\widehat{\beta} - \beta}{\sqrt{\frac{n\widehat{\sigma^2}}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$

has a $t$ distribution with $n - 2$ degrees of freedom. Therefore,

$$P\left[-t_{\gamma/2}(n-2) \le \frac{\widehat{\beta} - \beta}{\sqrt{\frac{n\widehat{\sigma^2}}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \le t_{\gamma/2}(n-2)\right] = 1 - \gamma,$$

and it follows that

$$\left[\widehat{\beta} - t_{\gamma/2}(n-2)\sqrt{\frac{n\widehat{\sigma^2}}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}\,,\ \widehat{\beta} + t_{\gamma/2}(n-2)\sqrt{\frac{n\widehat{\sigma^2}}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]$$

is a $100(1 - \gamma)\%$ confidence interval for $\beta$.

Similarly,

$$T_2 = \frac{\frac{\sqrt{n}(\widehat{\alpha} - \alpha)}{\sigma}}{\sqrt{\frac{n\widehat{\sigma^2}}{\sigma^2(n-2)}}} = \frac{\widehat{\alpha} - \alpha}{\sqrt{\frac{\widehat{\sigma^2}}{n-2}}}$$

has a $t$ distribution with $n - 2$ degrees of freedom. Thus, $T_2$ can be used to make inferences about $\alpha$. (See Exercise 7.6-14.) The fact that $n\widehat{\sigma^2}/\sigma^2$ has a chi-square distribution with $n-2$ degrees of freedom can be used to make inferences about the variance $\sigma^2$. (See Exercise 7.6-15.)

We have noted that $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}(x - \bar{x})$ is a point estimate for the mean of $Y$ for some given $x$, or we could think of this as a prediction of the value of $Y$ for this given $x$. But how close is $\widehat{Y}$ to the mean of $Y$ or to $Y$ itself? We shall now find a confidence interval for $\alpha + \beta(x - \bar{x})$ and a prediction interval for $Y$, given a particular value of $x$.

To find a confidence interval for

$$E(Y) = \mu(x) = \alpha + \beta(x - \bar{x}),$$

let

$$\widehat{Y} = \widehat{\alpha} + \widehat{\beta}(x - \bar{x}).$$

Recall that $\widehat{Y}$ is a linear combination of normally and independently distributed random variables $\widehat{\alpha}$ and $\widehat{\beta}$, so $\widehat{Y}$ has a normal distribution. Furthermore,

$$\begin{aligned} E(\widehat{Y}) &= E[\widehat{\alpha} + \widehat{\beta}(x - \bar{x})] \\ &= \alpha + \beta(x - \bar{x}) \end{aligned}$$

and

$$\begin{aligned}\operatorname{Var}(\widehat{Y}) &= \operatorname{Var}[\widehat{\alpha} + \widehat{\beta}\,(x - \bar{x}\,)] \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x}\,)^2}(x - \bar{x}\,)^2 \\ &= \sigma^2\left[\frac{1}{n} + \frac{(x - \bar{x}\,)^2}{\sum_{i=1}^{n}(x_i - \bar{x}\,)^2}\right].\end{aligned}$$

Recall that the distribution of $n\widehat{\sigma^2}/\sigma^2$ is $\chi^2(n-2)$. Since $\widehat{\alpha}$ and $\widehat{\beta}$ are independent of $\widehat{\sigma^2}$, we can form the $t$ statistic

$$T = \frac{\widehat{\alpha} + \widehat{\beta}(x - \bar{x}\,) - [\alpha + \beta(x - \bar{x}\,)]}{\sigma\sqrt{\dfrac{1}{n} + \dfrac{(x - \bar{x}\,)^2}{\sum_{i=1}^{n}(x_i - \bar{x}\,)^2}}} \Bigg/ \sqrt{\frac{n\widehat{\sigma^2}}{(n-2)\sigma^2}},$$

which has a $t$ distribution with $r = n-2$ degrees of freedom. Next we select $t_{\gamma/2}(n–2)$ from Table VI in Appendix B so that

$$P[-t_{\gamma/2}(n-2) \le T \le t_{\gamma/2}(n-2)] = 1 - \gamma.$$

This becomes

$$\begin{aligned}P[\widehat{\alpha} + \widehat{\beta}(x - \bar{x}\,) - ct_{\gamma/2}(n-2) &\le \alpha + \beta(x - \bar{x}\,) \\ &\le \widehat{\alpha} + \widehat{\beta}(x - \bar{x}\,) + ct_{\gamma/2}(n-2)] \\ &= 1 - \gamma,\end{aligned}$$

where

$$c = \sqrt{\frac{n\widehat{\sigma^2}}{n-2}}\sqrt{\frac{1}{n} + \frac{(x - \bar{x}\,)^2}{\sum_{i=1}^{n}(x_i - \bar{x}\,)^2}}.$$

Thus, the endpoints for a $100(1-\gamma)\%$ confidence interval for $\mu(x) = \alpha + \beta(x - \bar{x}\,)$ are

$$\widehat{\alpha} + \widehat{\beta}(x - \bar{x}\,) \pm ct_{\gamma/2}(n-2).$$

Note that the width of this interval depends on the particular value of $x$, because $c$ depends on $x$. (See Example 7.6-1.)

We have used $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ to estimate $\alpha$ and $\beta$. Suppose that we are given a value of $x$, say, $x_{n+1}$. A point estimate of the corresponding value of $Y$ is

$$\widehat{y}_{n+1} = \widehat{\alpha} + \widehat{\beta}(x_{n+1} - \bar{x}\,).$$

However, $\widehat{y}_{n+1}$ is just one possible value of the random variable

$$Y_{n+1} = \alpha + \beta(x_{n+1} - \bar{x}\,) + \varepsilon_{n+1}.$$

What can we say about possible values for $Y_{n+1}$? We shall now obtain a **prediction interval** for $Y_{n+1}$ when $x = x_{n+1}$ that is similar to the confidence interval for the mean of $Y$ when $x = x_{n+1}$.

We have

$$Y_{n+1} = \alpha + \beta(x_{n+1} - \bar{x}\,) + \varepsilon_{n+1},$$

where $\varepsilon_{n+1}$ is $N(0, \sigma^2)$. Now,

$$W = Y_{n+1} - \widehat{\alpha} - \widehat{\beta}\,(x_{n+1} - \bar{x}\,)$$

is a linear combination of normally and independently distributed random variables, so $W$ has a normal distribution. The mean of $W$ is

$$\begin{aligned} E(W) &= E[Y_{n+1} - \widehat{\alpha} - \widehat{\beta}\,(x_{n+1} - \bar{x}\,)] \\ &= \alpha + \beta(x_{n+1} - \bar{x}\,) - \alpha - \beta(x_{n+1} - \bar{x}\,) = 0. \end{aligned}$$

Since $Y_{n+1}, \widehat{\alpha}$ and $\widehat{\beta}$ are independent, the variance of $W$ is

$$\begin{aligned} \text{Var}(W) &= \sigma^2 + \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x}\,)^2}\,(x_{n+1} - \bar{x}\,)^2 \\ &= \sigma^2\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x}\,)^2}{\sum_{i=1}^{n}(x_i - \bar{x}\,)^2}\right]. \end{aligned}$$

Recall that $n\widehat{\sigma^2}/[(n-2)\sigma^2]$ is $\chi^2(n-2)$. Since $Y_{n+1}, \widehat{\alpha}$, and $\widehat{\beta}$ are independent of $\widehat{\sigma^2}$, we can form the $t$ statistic

$$T = \frac{Y_{n+1} - \widehat{\alpha} - \widehat{\beta}\,(x_{n+1} - \bar{x}\,)}{\sigma\sqrt{1 + \dfrac{1}{n} + \dfrac{(x_{n+1} - \bar{x}\,)^2}{\sum_{i=1}^{n}(x_i - \bar{x}\,)^2}}} \Bigg/ \sqrt{\frac{n\widehat{\sigma^2}}{(n-2)\sigma^2}},$$

which has a $t$ distribution with $r = n - 2$ degrees of freedom. Now we select a constant $t_{\gamma/2}(n-2)$ from Table VI in Appendix B so that

$$P[-t_{\gamma/2}(n-2) \le T \le t_{\gamma/2}(n-2)] = 1 - \gamma.$$

Solving this inequality for $Y_{n+1}$, we have

$$\begin{aligned} P[\widehat{\alpha} + \widehat{\beta}\,(x_{n+1} - \bar{x}\,) - d\,t_{\gamma/2}(n-2) &\le Y_{n+1} \\ &\le \widehat{\alpha} + \widehat{\beta}(x_{n+1} - \bar{x}\,) + dt_{\gamma/2}(n-2)] \\ &= 1 - \gamma, \end{aligned}$$

where

$$d = \sqrt{\frac{n\widehat{\sigma^2}}{n-2}}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x}\,)^2}{\sum_{i=1}^{n}(x_i - \bar{x}\,)^2}}.$$

Thus, the endpoints for a $100(1-\gamma)\%$ prediction interval for $Y_{n+1}$ are

$$\widehat{\alpha} + \widehat{\beta}(x_{n+1} - \bar{x}\,) \pm dt_{\gamma/2}(n-2).$$

Observe that

$$d^2 = c^2 + \frac{n\widehat{\sigma^2}}{n-2}$$

when $x_{n+1} = x$, implying that the $100(1-\gamma)\%$ prediction interval for $Y$ at $X = x$ is somewhat wider than the $100(1-\gamma)\%$ prediction interval for $\mu(x)$. This makes sense, since the difference between one observation of $Y$ (at a given $X$) and its predictor

tends to vary more than the difference between the mean of the entire population of $Y$ values (at the same $X$) and its estimator.

The collection of all $100(1-\gamma)\%$ confidence intervals for $\{\mu(x) : -\infty < x < \infty\}$ is called a **pointwise** $100(1-\gamma)\%$ **confidence band** for $\mu(x)$. Similarly, the collection of all $100(1-\gamma)\%$ prediction intervals for $\{Y(x) = \alpha + \beta x + \varepsilon : -\infty < x < \infty\}$ is called a **pointwise** $100(1-\gamma)\%$ **prediction band** for $Y$. Note, from the expressions for $c$ and $d$ in the confidence and prediction intervals, respectively, that these bands are narrowest at $x = \bar{x}$.

We shall now use the data in Example 6.5-1 to illustrate a 95% confidence interval for $\mu(x)$ and a 95% prediction interval for $Y$ for a given value of $x$. To find such intervals, we use Equations 6.5-1, 6.5-2, and 6.5-4.

**Example 7.6-1**

To find a 95% confidence interval for $\mu(x)$ using the data in Example 6.5-1, note that we have already found that $\bar{x} = 68.3$, $\widehat{\alpha} = 81.3$, $\widehat{\beta} = 561.1/756.1 = 0.7421$, and $\widehat{\sigma^2} = 21.7709$. We also need

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\right)\left(\sum_{i=1}^{n} x_i\right)^2$$
$$= 47,405 - \frac{683^2}{10} = 756.1.$$

For 95% confidence, $t_{0.025}(8) = 2.306$. When $x = 60$, the endpoints for a 95% confidence interval for $\mu(60)$ are

$$81.3 + 0.7421(60 - 68.3) \pm \left[\sqrt{\frac{10(21.7709)}{8}}\sqrt{\frac{1}{10} + \frac{(60-68.3)^2}{756.1}}\right](2.306),$$

or

$$75.1406 \pm 5.2589.$$

Similarly, when $x = 70$, the endpoints for a 95% confidence interval for $\mu(70)$ are

$$82.5616 \pm 3.8761.$$

Note that the lengths of these intervals depend on the particular value of $x$. A pointwise 95% confidence band for $\mu(x)$ is graphed in Figure 7.6-1(a) along with the scatter diagram and $\widehat{y} = \widehat{\alpha} + \widehat{\beta}(x - \bar{x})$.

The endpoints for a 95% prediction interval for $Y$ when $x = 60$ are

$$81.3 + 0.7421(60 - 68.3) \pm \left[\sqrt{\frac{10(21.7709)}{8}}\sqrt{1.1 + \frac{(60-68.3)^2}{756.1}}\right](2.306),$$

or

$$75.1406 \pm 13.1289.$$

Note that this interval is much wider than the confidence interval for $\mu(60)$. In Figure 7.6-1(b), the pointwise 95% prediction band for $Y$ is graphed along with the scatter diagram and the least squares regression line. ■

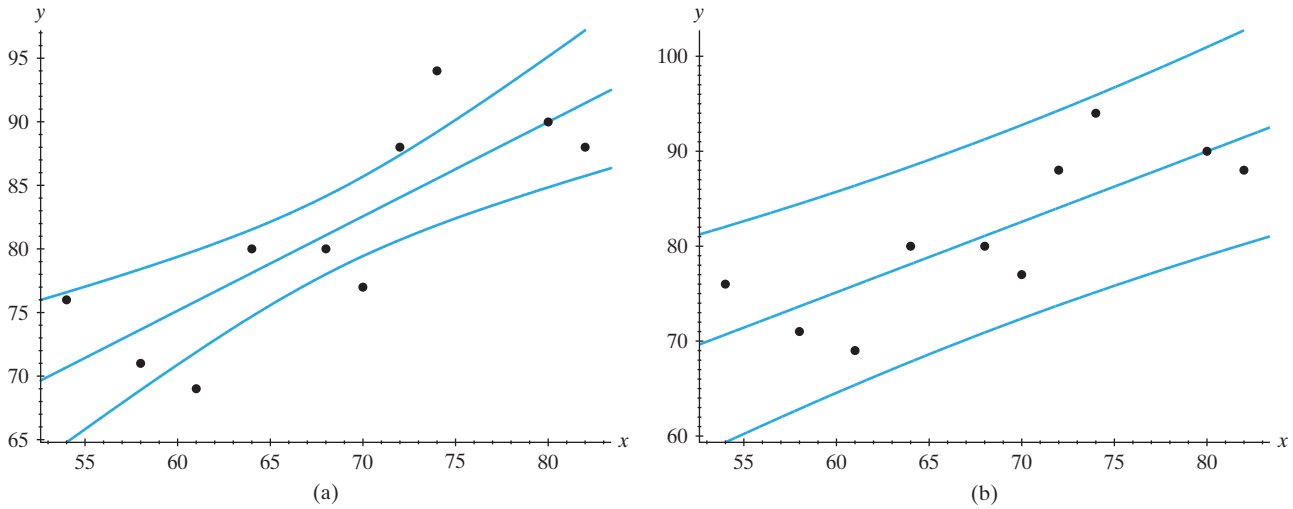



**Figure 7.6-1** A pointwise 95% (a) confidence band for $\mu(x)$ and (b) prediction band for $Y$

We now generalize the simple regression model to the **multiple regression** case. Suppose we observe several $x$-values—say, $x_1, x_2, \ldots, x_k$—along with the $y$-value. For example, suppose that $x_1$ equals the student's ACT composite score, $x_2$ equals the student's high school class rank, and $y$ equals the student's first-year GPA in college. We want to estimate a regression function $E(Y) = \mu(x_1, x_2, \ldots, x_k)$ from some observed data. If

$$\mu(x_1, x_2, \ldots, x_k) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

then we say that we have a **linear model** because this expression is linear in the coefficients $\beta_1, \beta_2, \ldots, \beta_k$.

To illustrate, note that the model in Section 6.5 is linear in $\alpha = \beta_1$ and $\beta = \beta_2$, with $x_1 = 1$ and $x_2 = x$, giving the mean $\alpha + \beta x$. (For convenience, there the mean of the $x$-values was subtracted from $x$.) Suppose, however, that we had wished to use the cubic function $\beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3$ as the mean. This cubic expression still provides a linear model (i.e., linear in the $\beta$-values), and we would take $x_1 = 1$, $x_2 = x$, $x_3 = x^2$, and $x_4 = x^3$.

Say our $n$ observation points are

$$(x_{1j}, x_{2j}, \ldots, x_{kj}, y_j), \qquad j = 1, 2, \ldots, n.$$

To fit the linear model $\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ by the method of least squares, we minimize

$$G = \sum_{j=1}^{n} (y_j - \beta_1 x_{1j} - \beta_2 x_{2j} - \cdots - \beta_k x_{kj})^2.$$

If we equate the $k$ first order partial derivatives

$$\frac{\partial G}{\partial \beta_i} = \sum_{j=1}^{n} (-2)(y_j - \beta_1 x_{1j} - \beta_2 x_{2j} - \cdots - \beta_k x_{kj})(x_{ij}), \qquad i = 1, 2, \ldots, k,$$

to zero, we obtain the $k$ **normal equations**

$$\begin{aligned}
\beta_1 \sum_{j=1}^{n} x_{1j}^2 &+ \beta_2 \sum_{j=1}^{n} x_{1j}x_{2j} + \cdots + \beta_k \sum_{j=1}^{n} x_{1j}x_{kj} = \sum_{j=1}^{n} x_{1j}y_j,\\
\beta_1 \sum_{j=1}^{n} x_{2j}x_{1j} &+ \beta_2 \sum_{j=1}^{n} x_{2j}^2 + \cdots + \beta_k \sum_{j=1}^{n} x_{2j}x_{kj} = \sum_{j=1}^{n} x_{2j}y_j,\\
&\vdots \qquad\qquad \vdots \qquad\quad \ddots \qquad \vdots \qquad\qquad \vdots\\
\beta_1 \sum_{j=1}^{n} x_{kj}x_{1j} &+ \beta_2 \sum_{j=1}^{n} x_{kj}x_{2j} + \cdots + \beta_k \sum_{j=1}^{n} x_{kj}^2 = \sum_{j=1}^{n} x_{kj}y_j.
\end{aligned}$$

The solution of the preceding $k$ equations provides the least squares estimates of $\beta_1, \beta_2, \ldots, \beta_k$. These estimates are also maximum likelihood estimates of $\beta_1, \beta_2, \ldots, \beta_k$, provided that the random variables $Y_1, Y_2, \ldots, Y_n$ are mutually independent and $Y_j$ is $N(\beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj}, \sigma^2)$, $j = 1, 2, \ldots, n$.

**Example 7.6-2** By the method of least squares, we fit $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ to the five observed points $(x_1, x_2, x_3, y)$:

$$(1,1,0,4), \quad (1,0,1,3), \quad (1,2,3,2), \quad (1,3,0,6), \quad (1,0,0,1).$$

Note that $x_1 = 1$ in each point, so we are really fitting $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3$. Since

$$\begin{aligned}
&\sum_{j=1}^{5} x_{1j}^2 = 5, \quad \sum_{j=1}^{5} x_{1j}x_{2j} = 6, \quad \sum_{j=1}^{5} x_{1j}x_{3j} = 4, \quad \sum_{j=1}^{5} x_{1j}y_j = 16,\\
&\sum_{j=1}^{5} x_{2j}x_{1j} = 6, \quad \sum_{j=1}^{5} x_{2j}^2 = 14, \quad \sum_{j=1}^{5} x_{2j}x_{3j} = 6, \quad \sum_{j=1}^{5} x_{2j}y_j = 26,\\
&\sum_{j=1}^{5} x_{3j}x_{1j} = 4, \quad \sum_{j=1}^{5} x_{3j}x_{2j} = 6, \quad \sum_{j=1}^{5} x_{3j}^2 = 10, \quad \sum_{j=1}^{5} x_{3j}y_j = 9,
\end{aligned}$$

the normal equations are

$$\begin{aligned}
5\beta_1 + 6\beta_2 + 4\beta_3 &= 16,\\
6\beta_1 + 14\beta_2 + 6\beta_3 &= 26,\\
4\beta_1 + 6\beta_2 + 10\beta_3 &= 9.
\end{aligned}$$

Solving these three linear equations in three unknowns, we obtain

$$\widehat{\beta}_1 = \frac{274}{112}, \qquad \widehat{\beta}_2 = \frac{127}{112}, \qquad \widehat{\beta}_3 = -\frac{85}{112}.$$

Thus, the least squares fit is

$$y = \frac{274x_1 + 127x_2 - 85x_3}{112}.$$

If $x_1$ always equals 1, then the equation reads

$$y = \frac{274 + 127x_2 - 85x_3}{112}.$$

It is interesting to observe that the usual two-sample problem is actually a linear model. Let $\beta_1 = \mu_1$ and $\beta_2 = \mu_2$, and consider $n$ pairs of $(x_1, x_2)$ that equal $(1, 0)$ and $m$ pairs that equal $(0, 1)$. This would require each of the first $n$ variables $Y_1, Y_2, \ldots, Y_n$ to have the mean

$$\beta_1 \cdot 1 + \beta_2 \cdot 0 = \beta_1 = \mu_1$$

and the next $m$ variables $Y_{n+1}, Y_{n+2}, \ldots, Y_{n+m}$ to have the mean

$$\beta_1 \cdot 0 + \beta_2 \cdot 1 = \beta_2 = \mu_2.$$

This is the background of the two-sample problem, but with the usual $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ replaced by $Y_1, Y_2, \ldots, Y_n$ and $Y_{n+1}, Y_{n+2}, \ldots, Y_{n+m}$, respectively.

## Exercises

**7.6-1.** The mean of $Y$ when $x = 0$ in the simple linear regression model is $\alpha - \beta\bar{x} = \alpha_1$. The least squares estimator of $\alpha_1$ is $\widehat{\alpha} - \widehat{\beta}\bar{x} = \widehat{\alpha}_1$.

**(a)** Find the distribution of $\widehat{\alpha}_1$ under the usual model assumptions.

**(b)** Obtain an expression for a $100(1 - \gamma)\%$ two-sided confidence interval for $\alpha_1$.

**7.6-2.** Obtain a two-sided $100(1 - \gamma)\%$ prediction interval for the average of $m$ future independent observations taken at the same $X$-value, $x^*$.

**7.6-3.** For the data given in Exercise 6.5-3, with the usual assumptions,

**(a)** Find a 95% confidence interval for $\mu(x)$ when $x = 68, 75$, and 82.

**(b)** Find a 95% prediction interval for $Y$ when $x = 68, 75$, and 82.

**7.6-4.** For the data given in Exercise 6.5-4, with the usual assumptions,

**(a)** Find a 95% confidence interval for $\mu(x)$ when $x = 2, 3$, and 4.

**(b)** Find a 95% prediction interval for $Y$ when $x = 2, 3$, and 4.

**7.6-5.** For the cigarette data in Exercise 6.5-7, with the usual assumptions,

**(a)** Find a 95% confidence interval for $\mu(x)$ when $x = 5, 10$, and 15.

**(b)** Determine a 95% prediction interval for $Y$ when $x = 5, 10$, and 15.

**7.6-6.** A computer center recorded the number of programs it maintained during each of 10 consecutive years.

**(a)** Calculate the least squares regression line for the data shown.

**(b)** Plot the points and the line on the same graph.

**(c)** Find a 95% prediction interval for the number of programs in year 11 under the usual assumptions.

| Year | Number of Programs |
|---|---|
| 1 | 430 |
| 2 | 480 |
| 3 | 565 |
| 4 | 790 |
| 5 | 885 |
| 6 | 960 |
| 7 | 1200 |
| 8 | 1380 |
| 9 | 1530 |
| 10 | 1591 |

**7.6-7.** For the ACT scores in Exercise 6.5-6, with the usual assumptions,

**(a)** Find a 95% confidence interval for $\mu(x)$ when $x = 17, 20, 23, 26$, and 29.

**(b)** Determine a 90% prediction interval for $Y$ when $x = 17, 20, 23, 26$, and 29.

**7.6-8.** By the method of least squares, fit the regression plane $y = \beta_1 + \beta_2 x_1 + \beta_3 x_2$ to the following 12 observations of $(x_1, x_2, y)$: (1, 1, 6), (0, 2, 3), (3, 0, 10),

(–2, 0, –4), (–1, 2, 0), (0, 0, 1), (2, 1, 8), (–1, –1, –2), (0, –3, –3), (2, 1, 5), (1, 1, 1), (–1, 0, –2).

**7.6-9.** By the method of least squares, fit the cubic equation $y = \beta_1+\beta_2x+\beta_3x^2+\beta_4x^3$ to the following 10 observed data points $(x, y)$: (0, 1), (–1, –3), (0, 3), (1, 3), (–1, –1), (2, 10), (0, 0), (–2, –9), (–1, –2), (2, 8).

**7.6-10.** We would like to fit the quadratic curve $y = \beta_1 + \beta_2x + \beta_3x^2$ to a set of points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ by the method of least squares. To do this, let

$$h(\beta_1, \beta_2, \beta_3) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2x_i - \beta_3x_i^2)^2.$$

**(a)** By setting the three first partial derivatives of $h$ with respect to $\beta_1$, $\beta_2$, and $\beta_3$ equal to 0, show that $\beta_1$, $\beta_2$, and $\beta_3$ satisfy the following set of equations (called normal equations), all of which are sums going from 1 to $n$:

$$\beta_1 n + \beta_2 \sum x_i + \beta_3 \sum x_i^2 = \sum y_i;$$

$$\beta_1 \sum x_i + \beta_2 \sum x_i^2 + \beta_3 \sum x_i^3 = \sum x_i y_i;$$

$$\beta_1 \sum x_i^2 + \beta_2 \sum x_i^3 + \beta_3 \sum x_i^4 = \sum x_i^2 y_i.$$

**(b)** For the data

(6.91, 17.52) (4.32, 22.69) (2.38, 17.61) (7.98, 14.29)
(8.26, 10.77) (2.00, 12.87) (3.10, 18.63) (7.69, 16.77)
(2.21, 14.97) (3.42, 19.16) (8.18, 11.15) (5.39, 22.41)
(1.19, 7.50) (3.21, 19.06) (5.47, 23.89) (7.35, 16.63)
(2.32, 15.09) (7.54, 14.75) (1.27, 10.75) (7.33,17.42)
(8.41, 9.40) (8.72, 9.83) (6.09, 22.33) (5.30, 21.37)
(7.30, 17.36)

$n = 25$, $\sum x_i = 133.34$, $\sum x_i^2 = 867.75$, $\sum x_i^3 = 6197.21$, $\sum x_i^4 = 46{,}318.88$, $\sum y_i = 404.22$, $\sum x_iy_i = 2138.38$, and $\sum x_i^2y_i = 13{,}380.30$. Show that $a = -1.88$, $b = 9.86$, and $c = -0.995$.

**(c)** Plot the points and the linear regression line for these data.

**(d)** Calculate and plot the residuals. Does linear regression seem to be appropriate?

**(e)** Show that the least squares quadratic regression line is $\widehat{y} = -1.88 + 9.86x - 0.995x^2$.

**(f)** Plot the points and this least squares quadratic regression curve on the same graph.

**(g)** Plot the residuals for quadratic regression and compare this plot with that in part (d).

**7.6-11.** (The information presented in this exercise comes from the Westview Blueberry Farm and National Oceanic and Atmospheric Administration Reports [NOAA].) For the following paired data, $(x, y)$, $x$ gives the Holland, Michigan, rainfall for June, and $y$ gives the blueberry production in thousands of pounds from the Westview Blueberry Farm:

| | | | |
|---|---|---|---|
| (4.11, 56.2) | (5.49, 45.3) | (5.35, 31.0) | (6.53, 30.1) |
| (5.18, 40.0) | (4.89, 38.5) | (2.09, 50.0) | (1.40, 45.8) |
| (4.52, 45.9) | (1.11, 32.4) | (0.60, 18.2) | (3.80, 56.1) |

The data are from 1971 to 1989 for those years in which the last frost occurred May 10 or earlier.

**(a)** Find the correlation coefficient for these data.

**(b)** Find the least squares regression line.

**(c)** Make a scatter plot of the data with the least squares regression line on the plot.

**(d)** Calculate and plot the residuals. Does linear regression seem to be appropriate?

**(e)** Find the least squares quadratic regression curve.

**(f)** Calculate and plot the residuals. Does quadratic regression seem to be appropriate?

**(g)** Give a short interpretation of your results.

**7.6-12.** Explain why the model $\mu(x) = \beta_1e^{\beta_2x}$ is not a linear model. Would taking the logarithms of both sides yield a linear model for $\ln \mu(x)$?

**7.6-13.** Show that

$$\sum_{i=1}^{n} [Y_i - \alpha - \beta(x_i - \bar{x})]^2$$
$$= n(\widehat{\alpha} - \alpha)^2 + (\widehat{\beta} - \beta)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$
$$+ \sum_{i=1}^{n} [Y_i - \widehat{\alpha} - \widehat{\beta}(x_i - \bar{x})]^2.$$

**7.6-14.** Show that the endpoints for a $100(1 - \gamma)\%$ confidence interval for $\alpha$ are

$$\widehat{\alpha} \pm t_{\gamma/2}(n-2)\sqrt{\frac{\widehat{\sigma^2}}{n-2}}.$$

**7.6-15.** Show that a $100(1 - \gamma)\%$ confidence interval for $\sigma^2$ is

$$\left[\frac{n\widehat{\sigma^2}}{\chi^2_{\gamma/2}(n-2)}, \frac{n\widehat{\sigma^2}}{\chi^2_{1-\gamma/2}(n-2)}\right].$$

**7.6-16.** Find 95% confidence intervals for $\alpha$, $\beta$, and $\sigma^2$ for the predicted and earned grades data in Exercise 6.5-4.

**7.6-17.** Find 95% confidence intervals for $\alpha$, $\beta$, and $\sigma^2$ for the midterm and final exam scores data in Exercise 6.5-3.

**7.6-18.** Using the cigarette data in Exercise 6.5-7, find 95% confidence intervals for $\alpha$, $\beta$, and $\sigma^2$ under the usual assumptions.

**7.6-19.** Using the data in Exercise 6.5-8(a), find 95% confidence intervals for $\alpha$, $\beta$, and $\sigma^2$.

**7.6-20.** Using the ACT scores in Exercise 6.5-6, find 95% confidence intervals for $\alpha$, $\beta$, and $\sigma^2$ under the usual assumptions.

## 7.7* RESAMPLING METHODS

Sampling and resampling methods have become more useful in recent years due to the power of computers. These methods are even used in introductory courses to convince students that statistics have distributions—that is, that statistics are random variables with distributions. At this stage in the book, the reader should be convinced that this is true, although we did use some sampling in Section 5.6 to help sell the idea that the sample mean has an approximate normal distribution.

Resampling methods, however, are used for more than showing that statistics have certain distributions. Rather, they are needed in finding approximate distributions of certain statistics that are used to make statistical inferences. We already know a great deal about the distribution of $\overline{X}$, and resampling methods are not needed for $\overline{X}$. In particular, $\overline{X}$ has an approximate normal distribution with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$. Of course, if the latter is unknown, we can estimate it by $s/\sqrt{n}$ and note that $(\overline{X}-\mu)/(s/\sqrt{n})$ has an approximate $N(0,1)$ distribution, provided that the sample size is large enough and the underlying distribution is not too badly skewed with a long, heavy tail.

We know something about the distribution of $S^2$ *if the random sample arises from a normal distribution* or one fairly close to it. However, the statistic $S^2$ is not very robust, in that its distribution changes a great deal as the underlying distribution changes. It is not like $\overline{X}$, which always has an approximate normal distribution, provided that the mean $\mu$ and variance $\sigma^2$ of the underlying distribution exist. So what do we do about distributions of statistics like the sample variance $S^2$, whose distribution depends so much on having a given underlying distribution? We use resampling methods that essentially substitute computation for theory. We need to have some idea about the distributions of these various estimators to find confidence intervals for the corresponding parameters.

Let us now explain resampling. Suppose that we need to find the distribution of some statistic, such as $S^2$, but we do not believe that we are sampling from a normal distribution. We observe the values of $X_1, X_2, \ldots, X_n$ to be $x_1, x_2, \ldots, x_n$. Actually, if we know nothing about the underlying distribution, then the empirical distribution found by placing the weight $1/n$ on each $x_i$ is the best estimate of that distribution. Therefore, to get some idea about the distribution of $S^2$, let us take a random sample of size $n$ from this empirical distribution; then we are sampling from the $n$ values with replacement. We compute $S^2$ for that sample; say it is $s_1^2$. We then do it again, getting $s_2^2$. And again, we compute $s_3^2$. We continue to do this a large number of times, say, $N$, where $N$ might be 1000, 2000, or even 10,000. Once we have these $N$ values of $S^2$, we can construct a histogram, a stem-and-leaf display, or a $q$–$q$ plot—anything to help us get some information about the distribution of $S^2$ when the sample arises from this empirical distribution, which is an estimate of the real underlying distribution. Clearly, we must use the computer for all of this sampling. We illustrate the resampling procedure by using, not $S^2$, but a statistic called the *trimmed mean*.

Although we usually do not know the underlying distribution, we state that, in this illustration, it is of the Cauchy type, because there are certain basic ideas we want to review or introduce for the first time. The pdf of the Cauchy is

$$f(x) = \frac{1}{\pi(1+x^2)}, \qquad -\infty < x < \infty.$$

The cdf is

$$F(x) = \int_{-\infty}^{x} \frac{1}{\pi(1+w^2)}\,dw = \frac{1}{\pi}\arctan x + \frac{1}{2}, \qquad -\infty < x < \infty.$$

If we want to generate some $X$-values that have this distribution, we let $Y$ have the uniform distribution $U(0,1)$ and define $X$ by

$$Y = F(X) = \frac{1}{\pi}\arctan X + \frac{1}{2}$$

or, equivalently,

$$X = \tan\left[\pi\left(Y - \frac{1}{2}\right)\right].$$

We can generate 40 values of $Y$ on the computer and then calculate the 40 values of $X$. Let us now add $\theta = 5$ to each $X$-value to create a sample from a Cauchy distribution with a median of 5. That is, we have a random sample of 40 $W$-values, where $W = X + 5$. We will consider some statistics used to estimate the median, $\theta$, of this distribution. Of course, usually the value of the median is unknown, but here we know that it is equal to $\theta = 5$, and our statistics are estimates of this known number. These 40 values of $W$ are as follows, after ordering:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| −7.34 | −5.92 | −2.98 | 0.19 | 0.77 | 0.95 | 2.86 | 3.17 | 3.76 | 4.20 |
| 4.20 | 4.27 | 4.31 | 4.42 | 4.60 | 4.73 | 4.84 | 4.87 | 4.90 | 4.96 |
| 4.98 | 5.00 | 5.09 | 5.09 | 5.14 | 5.22 | 5.23 | 5.42 | 5.50 | 5.83 |
| 5.94 | 5.95 | 6.00 | 6.01 | 6.24 | 6.82 | 9.62 | 10.03 | 18.27 | 93.62 |

It is interesting to observe that many of these 40 values are between 3 and 7 and hence are close to $\theta = 5$; it is almost as if they had arisen from a normal distribution with mean $\mu = 5$ and $\sigma^2 = 1$. But then we note the outliers; these very large or small values occur because of the heavy and long tails of the Cauchy distribution and suggest that the sample mean $\overline{X}$ is not a very good estimator of the middle. And it is not in this sample, because $\bar{x} = 6.67$. In a more theoretical course, it can be shown that, due to the fact that the mean $\mu$ and the variance $\sigma^2$ do not exist for a Cauchy distribution, $\overline{X}$ is not any better than a single observation $X_i$ in estimating the median $\theta$. The sample median $\widetilde{m}$ is a much better estimate of $\theta$, as it is not influenced by the outliers. Here the median equals 4.97, which is fairly close to 5. Actually, the maximum likelihood estimator found by maximizing

$$L(\theta) = \prod_{i=1}^{40} \frac{1}{\pi[1+(x_i-\theta)^2]}$$

is extremely good but requires difficult numerical methods to compute. Then advanced theory shows that, in the case of a Cauchy distribution, a **trimmed mean**, found by ordering the sample, discarding the smallest and largest $3/8 = 37.5\%$ of the

sample, and averaging the middle 25%, is almost as good as the maximum likelihood estimator but is much easier to compute. This trimmed mean is usually denoted by $\overline{X}_{0.375}$; we use $\overline{X}_t$ for brevity, and here $\bar{x}_t = 4.96$. For this sample, it is not quite as good as the median; but, for most samples, it is better. Trimmed means are often very useful and many times are used with a smaller trimming percentage. For example, in sporting events such as skating and diving, often the smallest and largest of the judges' scores are discarded.

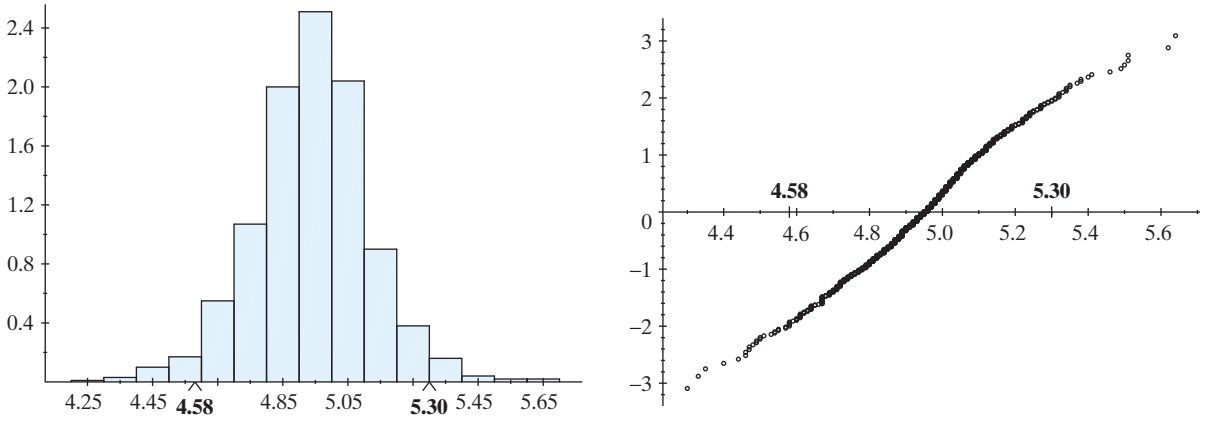
For this Cauchy example, let us resample from the empirical distribution created by placing the "probability" 1/40 on each of our 40 observations. With each of these samples, we find our trimmed mean $\overline{X}_t$. That is, we order the observations of each resample and average the middle 25% of the order statistics—namely, the middle 10 order statistics. We do this $N = 1000$ times, thus obtaining $N = 1000$ values of $\overline{X}_t$. These values are summarized with the histogram in Figure 7.7-1(a).

From this resampling procedure, which is called **bootstrapping**, we have some idea about the distribution if the sample arises from the empirical distribution and, hopefully, from the underlying distribution, which is approximated by the empirical distribution. While the distribution of the sample mean $\overline{X}$ is not normal if the sample arises from a Cauchy-type distribution, the approximate distribution of $\overline{X}_t$ is normal. From the histogram of trimmed mean values in Figure 7.7-1(a), that looks to be the case. This observation is supported by the $q$–$q$ plot in Figure 7.7-1(b) of the quantiles of a standard normal distribution versus those of the 1000 $\bar{x}_t$-values: The plot is very close to being a straight line.

How do we find a confidence interval for $\theta$? Recall that the middle of the distribution of $\overline{X}_t - \theta$ is zero. So a guess at $\theta$ would be the amount needed to move the histogram of $\overline{X}_t$-values over so that zero is more or less in the middle of the translated histogram. We recognize that this histogram was generated from the original sample $X_1, X_2, \ldots, X_{40}$ and thus is really only an estimate of the distribution of $\overline{X}_t$.

We could get a point estimate of $\theta$ by moving it over until its median (or mean) is at zero. Clearly, however, some error is incurred in doing so—and we really want some bounds for $\theta$ as given by a confidence interval.

To find that confidence interval, let us proceed as follows: In the $N = 1000$ resampled values of $\overline{X}_t$, we find two points—say, $c$ and $d$—such that about 25 values are less than $c$ and about 25 are greater than $d$. That is, $c$ and $d$ are about on




(a) Histogram of the trimmed means

(b) $N(0,1)$ versus trimmed-mean quantiles

**Figure 7.7-1** $N = 1000$ observations of trimmed means

the respective 2.5th and 97.5th percentiles of the empirical distribution of these $N = 1000$ resampled $\overline{X}_t$-values. Thus, $\theta$ should be big enough so that over 2.5% of the $\overline{X}_t$-values are less than $c$ and small enough so that over 2.5% of the $\overline{X}_t$-values are greater than $d$. This requires that $c < \theta$ and $\theta < d$; thus, $[c, d]$ serves as an approximate 95% confidence interval for $\theta$ as found by the **percentile method**. With our bootstrapped distribution of $N = 1000$ $\overline{X}_t$-values, this 95% confidence interval for $\theta$ runs from 4.58 to 5.30, and these two points are marked on the histogram and the $q$–$q$ plot. Clearly, we could change this percentage to other values, such as 90%.

This percentile method, associated with the bootstrap method, is a nonparametric procedure, as we make no assumptions about the underlying distribution. It is interesting to compare the answer it produces with that obtained by using the order statistics $Y_1 < Y_2 < \cdots < Y_{40}$. If the sample arises from a continuous-type distribution, then, with the use of a calculator or computer, we have, when $\theta$ is the median,

$$P(Y_{14} < \theta < Y_{27}) = \sum_{k=14}^{26} \binom{40}{k}\left(\frac{1}{2}\right)^{40} = 0.9615.$$

(See Section 7.5.) Since, in our illustration, $Y_{14} = 4.42$ and $Y_{27} = 5.23$, the interval $[4.42,\ 5.23]$ is an approximate 96% confidence interval for $\theta$. Of course, $\theta = 5$ is included in each of the two confidence intervals. In this case, the bootstrap confidence interval is a little more symmetric about $\theta = 5$ and somewhat shorter, but it did require much more work.

We have now illustrated bootstrapping, which allows us to substitute computation for theory to make statistical inferences about characteristics of the underlying distribution. This method is becoming more important as we encounter complicated data sets that clearly do not satisfy certain underlying assumptions. For example, consider the distribution of $T = (\overline{X} - \mu)/(S/\sqrt{n})$ when the random sample arises from an exponential distribution that has pdf $f(x) = e^{-x}$, $0 < x < \infty$, with mean $\mu = 1$. First, we will *not* use resampling, but we will simulate the distribution of $T$ when the sample size $n = 16$ by taking $N = 1000$ random samples from this known exponential distribution. Here

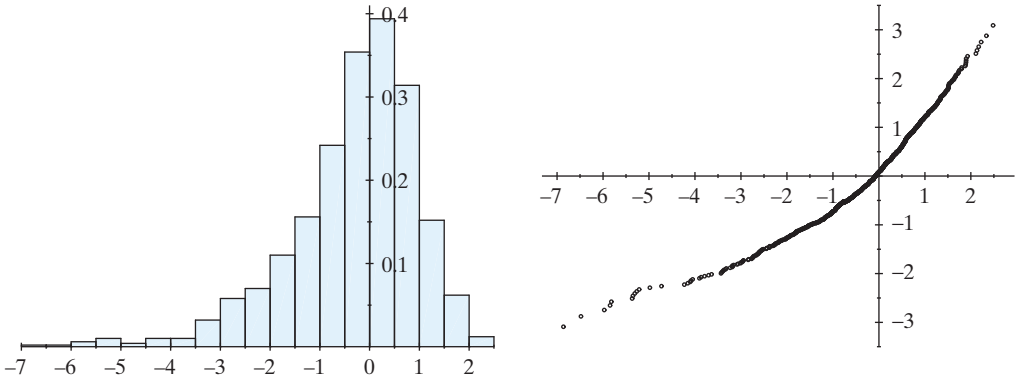
$$F(x) = \int_0^x e^{-w}\,dw = 1 - e^{-x}, \qquad 0 < x < \infty.$$

So $Y = F(X)$ means

$$X = -\ln(1 - Y)$$

and $X$ has that given exponential distribution with $\mu = 1$, provided that $Y$ has the uniform distribution $U(0, 1)$. With the computer, we select $n = 16$ values of $Y$, determine the corresponding $n = 16$ values of $X$, and, finally, compute the value of $T = (\overline{X} - 1)/(S/\sqrt{16})$—say, $T_1$. We repeat this process over and over again, obtaining not only $T_1$, but also the values of $T_2, T_3, \ldots, T_{1000}$. We have done this and display the histogram of the 1000 $T$-values in Figure 7.7-2(a). Moreover the $q$–$q$ plot with quantiles of $N(0, 1)$ on the $y$-axis is displayed in Figure 7.7-2(b). Both the histogram and the $q$–$q$ plot show that the distribution of $T$ in this case is skewed to the left.

In the preceding illustration, we knew the underlying distribution. Let us now sample from the exponential distribution with mean $\mu = 1$, but add a value $\theta$ to each $X$. Thus, we will try to estimate the new mean $\theta + 1$. The authors know the




(a) 1000 observations of $T$ (b) $N(0,1)$ quantiles versus $T$ quantiles

**Figure 7.7-2** $T$ observations from an exponential distribution

value of $\theta$, but the readers do not know it at this time. The observed 16 values of this random sample are

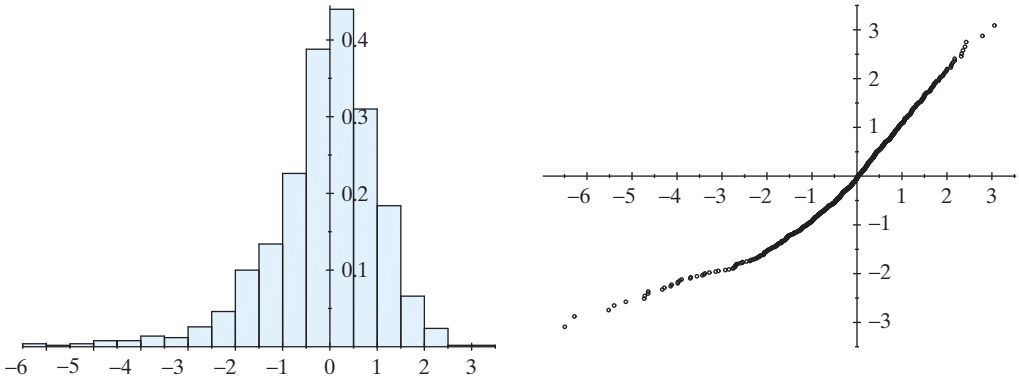
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 11.9776 | 9.3889 | 9.9798 | 13.4676 | 9.2895 | 10.1242 | 9.5798 | 9.3148 |
| 9.0605 | 9.1680 | 11.0394 | 9.1083 | 10.3720 | 9.0523 | 13.2969 | 10.5852 |

At this point we are trying to find a confidence interval for $\mu = \theta + 1$, and we pretend that we do not know that the underlying distribution is exponential. Actually, this is the case in practice: We do not know the underlying distribution. So we use the empirical distribution as the best guess of the underlying distribution; it is found by placing the weight 1/16 on each of the observations. The mean of this empirical distribution is $\bar{x} = 10.3003$. Therefore, we obtain some idea about the distribution of $T$ by now simulating

$$T = \frac{\overline{X} - 10.3003}{S/\sqrt{16}}$$

with $N = 1000$ random samples from the empirical distribution.

We obtain $t_1, t_2, \ldots, t_{1000}$, and these values are used to construct a histogram, shown in Figure 7.7-3(a), and a $q$–$q$ plot, illustrated in Figure 7.7-3(b). These two




(a) 1000 observations of $T$ (b) $N(0,1)$ quantiles versus $T$ quantiles

**Figure 7.7-3** $T$ observations from an empirical distribution

figures look somewhat like those in Figure 7.7-2. Moreover, the 0.025th and 0.975th quantiles of the 1000 $t$-values are $c = -3.1384$ and $d = 1.8167$, respectively.

Now we have some idea about the 2.5th and 97.5th percentiles of the $T$ distribution. Hence, as a very rough approximation, we can write

$$P\left(-3.1384 \leq \frac{\overline{X} - \mu}{S/\sqrt{16}} \leq 1.8167\right) \approx 0.95.$$

This formula leads to the rough approximate 95% confidence interval

$$[\,\bar{x} - 1.8167s/\sqrt{16},\ \bar{x} - (-3.1384)s/\sqrt{16}\,]$$

once the $\bar{x}$ and $s$ of the *original* sample are substituted. With $\bar{x} = 10.3003$ and $s = 1.4544$, we have

$$[10.3003 - 1.8167(1.4544)/4,\ 10.3003 + 3.1384(1.4544)/4] = [9.6397,\ 11.4414]$$

as a 95% approximate confidence interval for $\mu = \theta + 1$. Note that, because we added $\theta = 9$ to each $x$-value, the interval does cover $\theta + 1 = 10$.

It is easy to see how this procedure gets its name, because it is like "pulling yourself up by your own bootstraps," with the empirical distribution acting as the bootstraps.

## Exercises

**7.7-1.** If time and computing facilities are available, consider the following 40 losses, due to wind-related catastrophes, that were recorded to the nearest \$1 million (these data include only those losses of \$2 million or more, and, for convenience, they have been ordered and recorded in millions of dollars):

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 2 | 2 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 5 |
| 5 | 5 | 5 | 6 | 6 | 6 | 6 | 8 | 8 | 9 |
| 15 | 17 | 22 | 23 | 24 | 24 | 25 | 27 | 32 | 43 |

To illustrate bootstrapping, take resamples of size $n = 40$ as many as $N = 100$ times, computing the value of $T = (\overline{X} - 5)/(S/\sqrt{40})$ each time. Here the value 5 is the median of the original sample. Construct a histogram of the bootstrapped values of $T$.

**7.7-2.** Consider the following 16 observed values, rounded to the nearest tenth, from the exponential distribution that was given in this section:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 12.0 | 9.4 | 10.0 | 13.5 | 9.3 | 10.1 | 9.6 | 9.3 |
| 9.1 | 9.2 | 11.0 | 9.1 | 10.4 | 9.1 | 13.3 | 10.6 |

**(a)** Take resamples of size $n = 16$ from these observations about $N = 200$, times and compute $s^2$ each time. Construct a histogram of these 200 bootstrapped values of $S^2$.

**(b)** Simulate $N = 200$ random samples of size $n = 16$ from an exponential distribution with $\theta$ equal to the mean of the data in part (a) minus 9. For each sample, calculate the value of $s^2$. Construct a histogram of these 200 values of $S^2$.

**(c)** Construct a $q$–$q$ plot of the two sets of sample variances and compare these two empirical distributions of $S^2$.

**7.7-3.** Refer to the data in Example 7.5-1 and take resamples of size $n = 9$ exactly $N = 1000$ times and compute the fifth order statistic, $y_5$, each time.

**(a)** Construct a histogram of these $N = 1000$ fifth order statistics.

**(b)** Find a point estimate of the median, $\pi_{0.50}$.

**(c)** Also, calculate a 96% confidence interval for $\pi_{0.50}$ by finding two numbers, the first of which has $(1000)(0.02) = 20$ values less than it and the second has 20 values greater than it. How does this interval compare to the one given in that example?

**7.7-4.** Refer to the data in Example 7.5-2 and take resamples of size $n = 27$ exactly $N = 500$ times and compute the seventh order statistic, $y_7$, each time.

**(a)** Construct a histogram of these $N = 500$ seventh order statistics.

**(b)** Give a point estimate of $\pi_{0.25}$.

**(c)** Find an 82% confidence interval for $\pi_{0.25}$ by finding two numbers, the first of which has $(500)(0.09) = 45$ values less than it and the second has 205 values greater than it.

**(d)** How does this interval compare to the one given in that example?

**7.7-5.** Let $X_1, X_2, \ldots, X_{21}$ and $Y_1, Y_2, \ldots, X_{21}$ be independent random samples of sizes $n = 21$ and $m = 21$ from $N(0,1)$ distributions. Then $F = S_X^2/S_Y^2$ has an $F$ distribution with 20 and 20 degrees of freedom.

**(a)** Illustrate this situation empirically by simulating 100 observations of $F$.

**(i)** Plot a relative frequency histogram with the $F(20,20)$ pdf superimposed.

**(ii)** Construct a $q$–$q$ plot of the quantiles of $F(20,20)$ versus the order statistics of your simulated data. Is the plot linear?

**(b)** Consider the following 21 observations of the $N(0,1)$ random variable $X$:

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.1616 | −0.8593 | 0.3105 | 0.3932 | −0.2357 | 0.9697 | 1.3633 |
| −0.4166 | 0.7540 | −1.0570 | −0.1287 | −0.6172 | 0.3208 | 0.9637 |
| 0.2494 | −1.1907 | −2.4699 | −0.1931 | 1.2274 | −1.2826 | −1.1532 |

Consider also the following 21 observations of the $N(0,1)$ random variable $Y$:

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.4419 | −0.2313 | 0.9233 | −0.1203 | 1.7659 | −0.2022 | 0.9036 |
| −0.4996 | −0.8778 | −0.8574 | 2.7574 | 1.1033 | 0.7066 | 1.3595 |
| −0.0056 | −0.5545 | −0.1491 | −0.9774 | −0.0868 | 1.7462 | −0.2636 |

Sampling with replacement, resample with a sample of size 21 from each of these sets of observations. Calculate the value of $w = s_x^2/s_y^2$. Repeat in order to simulate 100 observations of $W$ from these two empirical distributions. Use the same graphical comparisons that you used in part (a) to see if the 100 observations represent observations from an approximate $F(20,20)$ distribution.

**(c)** Consider the following 21 observations of the exponential random variable $X$ with mean 1:

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.6958 | 1.6394 | 0.2464 | 1.5827 | 0.0201 | 0.4544 | 0.8427 |
| 0.6385 | 0.1307 | 1.0223 | 1.3423 | 1.6653 | 0.0081 | 5.2150 |
| 0.5453 | 0.08440 | 1.2346 | 0.5721 | 1.5167 | 0.4843 | 0.9145 |

Consider also the following 21 observations of the exponential random variable $Y$ with mean 1:

| | | | | | | |
|---|---|---|---|---|---|---|
| 1.1921 | 0.3708 | 0.0874 | 0.5696 | 0.1192 | 0.0164 | 1.6482 |
| 0.2453 | 0.4522 | 3.2312 | 1.4745 | 0.8870 | 2.8097 | 0.8533 |
| 0.1466 | 0.9494 | 0.0485 | 4.4379 | 1.1244 | 0.2624 | 1.3655 |

Sampling with replacement, resample with a sample of size 21 from each of these sets of observations. Calculate the value of $w = s_x^2/s_y^2$. Repeat in order to simulate 100 observations of $W$ from these two empirical distributions. Use the same graphical comparisons that you used in part (a) to see if the 100 observations represent observations from an approximate $F(20,20)$ distribution.

**7.7-6.** The following 54 pairs of data give, for Old Faithful geyser, the duration in minutes of an eruption and the time in minutes until the next eruption:

| | | | | | |
|---|---|---|---|---|---|
| (2.500, 72) | (4.467, 88) | (2.333, 62) | (5.000, 87) | (1.683, 57) | (4.500, 94) |
| (4.500, 91) | (2.083, 51) | (4.367, 98) | (1.583, 59) | (4.500, 93) | (4.550, 86) |
| (1.733, 70) | (2.150, 63) | (4.400, 91) | (3.983, 82) | (1.767, 58) | (4.317, 97) |
| (1.917, 59) | (4.583, 90) | (1.833, 58) | (4.767, 98) | (1.917, 55) | (4.433, 107) |
| (1.750, 61) | (4.583, 82) | (3.767, 91) | (1.833, 65) | (4.817, 97) | (1.900, 52) |
| (4.517, 94) | (2.000, 60) | (4.650, 84) | (1.817, 63) | (4.917, 91) | (4.000, 83) |
| (4.317, 84) | (2.133, 71) | (4.783, 83) | (4.217, 70) | (4.733, 81) | (2.000, 60) |
| (4.717, 91) | (1.917, 51) | (4.233, 85) | (1.567, 55) | (4.567, 98) | (2.133, 49) |
| (4.500, 85) | (1.717, 65) | (4.783, 102) | (1.850, 56) | (4.583, 86) | (1.733, 62) |

**(a)** Calculate the correlation coefficient, and construct a scatterplot, of these data.

**(b)** To estimate the distribution of the correlation coefficient, $R$, resample 500 samples of size 54 from the empirical distribution, and for each sample, calculate the value of $R$.

**(c)** Construct a histogram of these 500 observations of $R$.

**(d)** Simulate 500 samples of size 54 from a bivariate normal distribution with correlation coefficient equal to the correlation coefficient of the geyser data. For each sample of 54, calculate the correlation coefficient.

**(e)** Construct a histogram of the 500 observations of the correlation coefficient.

**(f)** Construct a $q$–$q$ plot of the 500 observations of $R$ from the bivariate normal distribution of part (d) versus the 500 observations in part (b). Do the two distributions of $R$ appear to be about equal?

HISTORICAL COMMENTS One topic among many important ones in this chapter is regression, a technique that leads to a mathematical model of the result of some process in terms of some associated (explanatory) variables. We create such models to give us some idea of the value of a response variable if we know the values of certain explanatory variables. If we have an idea of the form of the equation relating these variables, then we can "fit" this model to the data; that is, we can determine approximate values for the unknown parameters in the model from the data. Now, no model is exactly correct; but, as the well-known statistician George Box observed, "Some are useful." That is, while models may be wrong and we should check them as best we can, they may be good enough approximations to shed some light on the issues of interest.

Once satisfactory models are found, they may be used

1. to determine the effect of each explanatory variable (some may have very little effect and can be dropped),
2. to estimate the response variable for given values of important explanatory variables,
3. to predict the future, such as upcoming sales (although this sometimes should be done with great care),
4. to often substitute a cheaper explanatory variable for an expensive one that is difficult to obtain [such as chemical oxygen demand (COD) for biological oxygen demand (BOD)].

The name *bootstrap* and the resulting technique were first used by Brad Efron of Stanford University. Efron knew that the expression "to pull oneself up by his or her own bootstraps" seems to come from *The Surprising Adventures of Baron Munchausen* by Rudolph Erich Raspe. The baron had fallen from the sky and found himself in a hole 9 fathoms deep and had no idea how to get out. He comments as follows: "Looking down I observed that I had on a pair of boots with exceptionally sturdy straps. Grasping them firmly, I pulled with all my might. Soon I had hoisted myself to the top and stepped out on terra firma without further ado."

Of course, in statistical *bootstrapping*, statisticians pull themselves up by their bootstraps (the empirical distributions) by recognizing that the empirical distribution is the best estimate of the underlying distribution without a lot of other assumptions. So they use the empirical distribution as if it is the underlying distribution to find approximate distributions of statistics of interest.